Random variables and probability distributions (2)

MAE301 Applied Experimental Statistics

Yi Ren, Yabin Liao

School for Engineering of Matter, Transport Energy
Arizona State University

September 8, 2015

Outline

continuous random variables

Continuous random variable: can take real values

Probability density function (pdf) $f_X(x)$ of random variable X describes the probability:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x) dx \tag{1}$$

pdf properties:

$$f_X(x) \geq 0, \qquad \int_{-\infty}^{\infty} f_X(x) = 1 \tag{2}$$

mean and variance

Let X be a continuous random variable with pdf $f_X(x)$. The mean (expected value) of X is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f_X(x) dx. \tag{3}$$

The variance is

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx. \tag{4}$$

normal distribution

A normal distribution has pdf

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

Check if μ and σ are the mean and variance of the normal distribution.

When $\mu = 0$ and $\sigma = 1$, we have a **standard** normal distribution.

derivation of normal pdf

Consider throwing a dart at the origin of an x-y plane. You are aiming at the origin, but random errors in your throw will produce varying results. We assume that:

- errors in x and y directions are independent
- chance to hit anywhere on a circle is the same
- large errors are less likely than small errors

probability calculation under normal pdf

For a general normal distribution random variable $X \sim N(\mu, \sigma^2)$, the probability for X to assume a value between x_1 and x_2 can be calculated by using definition:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{6}$$

This integral does not have a closed-form solution.

cumulative distribution function (cdf)

The cdf for a random variable X is

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx. \tag{7}$$

Therefore

$$P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1) \tag{8}$$

transformation to standard normal

A general normal random variable $X \sim N(\mu, \sigma^2)$ can be transformed in to a standard normal random variable Z by

$$Z = \frac{X - \mu}{\sigma} \tag{9}$$

Probability calculation

$$P(x_1 \leq X \leq x_2) = P(\frac{x_1 - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{x_2 - \mu}{\sigma}) = P(z_1 \leq Z \leq z_2). \tag{10}$$

exercise

A certain type of storage battery lasts, on average, 3 years with a standard deviation of 0.5 years. Assuming that the battery life are normally distributed.

Determine the probability that a given battery will last more than 2.3 years.

Determine the probability that a given battery will last more than 2 but less than 3.5 years.

iid random variables

Let repeated measurements $x_1, x_2, \cdots, x_n$ be drawn from the same distribution. We can consider these measurements as realizations of n **identically and independently distributed** (iid) random variables:

$$X_1, \cdots, X_n \sim f_X(x), \tag{11}$$

with mean μ and variance σ^2.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the average of these measurements. The mean of $\bar{X}$ is

$$\mu_{\bar{X}} = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mu. \tag{12}$$

iid random variables (cont.)

The variance is

$$\sigma_{\bar{X}}^2 = E\left((\frac{1}{n}\sum_{i=1}^{n} X_i - \mu)^2 \right) = \frac{\sigma^2}{n}. \quad (13)$$

See derivation from discrete random variable.

Therefore, the average of normal random variables is a random variable:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}). \quad (14)$$

exercise

A voltage measurement X has a normal distribution with $\mu = 40$ (V) and $\sigma = 6$ (V).

Find the value of x such that $P(X \leq x) = 45\%$

Find the value of x such that $P(X \geq x) = 14\%$

Find the value of d_1 such that $P(\mu - d_1 \leq X \leq \mu + d_1) = 90\%$

If 3 measurements are made and averaged, find the value of d_2 such that $P(\mu - d_2 \leq X_{avg} \leq \mu + d_2) = 90\%$

exponential distribution

The exponential distribution is useful for modeling time to failure of component parts, or waiting time between events. It has probability density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

What are the mean and variance?

If on average 3 samples fail per hour during a fatigue test, determine the probability that the next failure occurs within 5 minutes (Ans.: 0.2212)

central limit theorem

Central limit theorem: If $\bar{X}$ is the mean of a random variable of size n taken from a population with mean μ and variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{16}$$

as $n \to \infty$, is the standard normal distribution $N(0, 1)$.

The sample size $n = 30$ is a guideline to use for the central limit theorem. The normal approximation will generally be good if $n \geq 30$. If $n < 30$, the approximation is good only if the population is not too different from a normal distribution.

Summary of the class

- Continuous random variable: probability density function, cumulative distribution function
- (population) mean and variance, sample mean and variance (are random variables!)
- normal distribution
- exponential distribution
- central limit theorem

Python code for demos in the class

```
## normal distribution pdf
from scipy.stats import norm
from scipy import stats
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1, 1)
mean, var = norm.stats(moments='mv')
x = np.linspace(norm.ppf(0.0001),norm.ppf(0.9999), 100)
ax.plot(x, norm.pdf(x),'r-', lw=5, alpha=0.6, label='norm pdf')
r = norm.rvs(size=100000)
ax.hist(r, normed=True, histtype='stepfilled', alpha=0.2)
ax.legend(loc='best', frameon=False)
plt.show()

## exercises
from scipy.stats import norm
# the battery problem
1-norm.cdf((2.3-3)/0.5)
norm.cdf(1) - norm.cdf(-2)
# the voltage problem
6*norm.ppf(0.45)+40 # inverse cdf (or called percent point function)
norm.cdf(norm.ppf(0.45)) # just to double check if ppf works
6*norm.ppf(0.86)+40
6*norm.ppf(0.95)
6/np.sqrt(3)*norm.ppf(0.95)
```

Python code for demos in the class

```
## exponential distribution
from scipy.stats import expon
fig, ax = plt.subplots(1, 1)
x = np.linspace(expon.ppf(0.01),expon.ppf(0.99), 100)
ax.plot(x, expon.pdf(x,scale=10),'r-', lw=5, alpha=0.6, label='expon pdf')
# exercise on exponential distribution
expon.cdf(5, scale=20)

## central limit theorem
#### a discrete case
from scipy import stats
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
xbar = []
ss = []
ssn = []
xk = np.arange(4) # variable takes 0, 1, 2, 3
pk = (0.1, 0.2, 0.3, 0.4) # probability masses are 0.1, 0.2, 0.3, 0.4
custm = stats.rv_discrete(name='custm', values=(xk, pk))
# calculate mean and variance
mu = np.sum(pk*xk)
variance = np.sum((xk-mu)**2*pk)

for i in np.arange(10000):
    R = custm.rvs(size=100)
    # calculate sample mean and sample variance
    xbar += [np.sum(R)/float(R.size)]
    ss += [np.sum((R-xbar[i])**2)/float(R.size-1)]
    ssn += [np.sum((R-xbar[i])**2)/float(R.size)]
hist, bins = np.histogram(xbar, bins=np.arange(0,3,0.1))
width = 0.7 * (bins[1] - bins[0])
center = (bins[:-1] + bins[1:]) / 2
plt.bar(center, hist, align='center', width=width)
plt.show()
```

Python code for demos in the class

```
## central limit theorem
#### bernoulli experiments (binomial)
from scipy.stats import binom
n, p = 1000, 0.3
mean, var = binom.stats(n, p, moments='mv')
x = np.arange(binom.ppf(0.0001, n, p), binom.ppf(0.9999, n, p))
fig, ax = plt.subplots(1, 1)
ax.plot(x, binom.pmf(x, n, p), 'bo', ms=8, label='binom pmf')
ax.vlines(x, 0, binom.pmf(x, n, p), colors='b', lw=5, alpha=0.5)

#### exponential distribution
from scipy.stats import expon
xbar = []
for i in np.arange(10000):
    R = expon.rvs(scale = 1, size=1000)
    # calculate sample mean and sample variance
    xbar += [np.sum(R)/float(R.size)]
hist, bins = np.histogram(xbar, bins=np.arange(0.9,1.1,0.01))
width = 0.7 * (bins[1] - bins[0])
center = (bins[:-1] + bins[1:]) / 2
plt.bar(center, hist, align='center', width=width)
plt.show()
```