Random variables and probability distributions (1)

MAE301 Applied Experimental Statistics

Yi Ren, Yabin Liao

School for Engineering of Matter, Transport Energy
Arizona State University

September 7, 2015

Outline

Discrete random variables

Summary

Appendix

discrete random variables

Discrete random variable: can take distinct discrete values

Probability mass function: $f(x_i) = P(X = x_i)$

What are the properties of $f(x)$?

Cumulative distribution function:
$F(x) = P(X \leq x_i) = \sum_{x_i \leq x} f(x_i)$

$F(-\inf) =?$, $F(\inf) =?$

mean and variance

For a given random variable X with probability mass function $f(x)$ defined on the set $\{x_1, \cdots, x_n\}$:

Mean (expected value, expectation): $\mu := E(X) := \sum_{i=1}^{n} x_i f(x_i)$

Variance: $\sigma^2 := E\left((X-\mu)^2\right) = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)$

σ is called the **standard deviation** of X.

sample mean and sample variance

Note that the mean and variance of a random variable is usually unknown. What we can observe is the sample mean and sample variance. Given samples $x_1, \cdots, x_m$ drawn from $f(x)$, we have:

Sample mean (average): $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$

Sample variance : $s^2 := \frac{1}{\mathbf{m-1}} \sum_{i=1}^{m} (x_i - \bar{x})^2$

degree of freedom

Why $m-1$ in the denominator of sample variance? Because sample variance has a DOF of $m-1$. But what does that mean? Well it means that $x_i - \bar{x}$ are sampled from a $m-1$ dimensional space. What? Well you see $\sum_{i=1}^{m}(x_i - \bar{x}) = 0$.

More explicit explanation: $s^2 = \frac{1}{\mathbf{m-1}} \sum_{i=1}^{m}(x_i - \bar{x})^2$ is an **unbiased** estimation of σ^2, i.e., $E(s^2) = \sigma^2$. (Proof?)

exercise

For a class of n students, consider each student's final grade as a random variable X_i, with mean μ_i and variance σ_i^2. What is the average grade of the class?

The average grade is defined as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Since X_i is a random variable, $\bar{X}$ is a random variable as well. The mean of $\bar{X}$ is:

$$\mu_X = E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_i. \quad (1)$$

exercise (cont.)

The variance of $\bar{X}$ is:

$$\begin{aligned}
\sigma_X^2 &= E\left((X-\mu_X)^2\right) \\
&= E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\sum_{i=1}^{n}\mu_i}{n}\right)^2\right) \\
&= E\left(\frac{\left(\sum_{i=1}^{n}(X_i-\mu_i)\right)^2}{n^2}\right) \\
&= \frac{\sum_{i=1}^{n}\sigma_i^2}{n^2}
\end{aligned} \tag{2}$$

What do you learn from here?

exercise

Assume that the total power output of a power plant can be mathematically modeled as $Z = 3X - 2Y$, where X and Y are two independent random variables: X takes values 1, 2 and 3 with probabilities 0.2, 0.3 and 0.5 respectively. Y takes values 3 and 5 with probabilities 0.5 and 0.5.

What are the mean and variance of Z?

binomial variable

A random variable is binomial when it describes the number of successes in a sequence of n independent success/failure experiments, each of which yields success with probability p.

E.g., the number of heads in 100 coin flips is a binomial variable.

The success/failure experiment is called a **Bernoulli experiment** or Bernoulli trial.

Bernoulli process

A **Bernoulli process** consists of repeated Bernoulli experiments.

- Each trial results in an outcome that may be classified as a success or a failure
- The probability of success, denoted by p, remains constant from trial to trial
- The repeated trials are independent

binomial distribution

Consider a Bernoulli process with n experiments, each has probability p to be successful. Let X be the number of successes in total. What values can X take and what are the probabilities?

Assume out of the n experiments, there are $m < n$ successes. There are in total $\binom{n}{m}$ combinations. The chance for each combination to happen is $p^m(1-p)^{n-m}$ (why?). So in total, the chance of having m successes is $\binom{n}{m}p^m(1-p)^{n-m}$.

The binomial distribution:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x \leq n, x \in \mathbb{Z} \tag{3}$$

A binomial variable X has $\mu = np$ and $\sigma^2 = np(1-p)$ (why?).

exercise

Traffic engineers install 10 street lights with new bulbs. The probability that a bulb fails within 50,000 hours of operation is 0.25. Assume that each of the bulbs fails independently.

- Determine the probability that fewer than two of the bulbs will fail within 50,000 hours of operation. (0.2440)
- Determine the probability that no bulbs will have to be replaced within 50,000 hours. (0.0563)
- Determine the probability that more than four of the bulbs will need replacing within 50,000 hours. (0.078127)

Summary of the class

- Discrete random variable: probability mass function, cumulative distribution function
- (population) mean and variance, sample mean and variance (are random variables!)
- binomial variable

Python code for the birthday problem

```
####################### Discrete random variable #######################
## mean, variance, sample mean, and sample variance
# define a distribution
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=10)
    # 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=10)
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

```
######################## Discrete random variable ########################
## histogram
discrete_uniform = np.random.randint(0,10,100000)
hist, bins = np.histogram(discrete_uniform, bins=10)
width = 0.7 * (bins[1] - bins[0])
center = (bins[:-1] + bins[1:]) / 2
plt.bar(center, hist, align='center', width=width)
plt.show()

## binomial distribution
from scipy.stats import binom
n, p = 20, 0.4
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)
```