Course Information

- Schedule and Homework
- Course Requirements
- Instructor's Information
- Links
- Main Page

Topics

1. Introduction
2. Summarizing Data: listing and grouping
3. Summarizing Data: Statistical Descriptions
4. Possibilities and Probabilities
5. Some rules of probability
6. Probability Distributions
7. Normal Distributions
8. Sampling Distributions
9. Problems of Estimation
10. Hypothesis Tests
11. Tests based on count data
12. Regression
13. Non-parametric Tests

Projects

1. Stock return distribution

- Forum

edit sec src prt

Site Manager

(Discrete) Probability Distributions

Random variables and distribution functions

A **random variable** is a quantity that takes values by chance (at random), for which a *distribution function* can be defined. They are often denoted by capital letters *X, Y* or *Z*.

A (cummulative) **distribution function** for a random variable X, is a function *F(x)* that takes real values and whose range is the interval [0,1]. The distribution function is defined by

$$F(x) := P(X \leq x), \tag{1}$$

that has the following properties:

- F(x) is non-decreasing: $F(x) \leq F(y)$ if $x < y$
- F(x) is continuous from above: $F(x+h) \to F(x)$ when $h > 0$ goes to 0.
- At the left end of the domain, F = 0, and at the right end of the domain, F=1.

Discrete vs. continuous random variables

A **discrete random** variable only takes either a finite or infinite (but countable) number of values.

Examples:

- The number of eggs that a hen lays in a given day (it can't be 2.3).
- The number of people going to a given soccer match.
- The number of students that come to class on a given day.
- The number of people in line at McDonalds on a given day and time.

A **continuous random** variable takes an infinite and uncountable number of values.

Examples:

- The amount of milk a cow produces on a given day (can be 1.7, 1.76 or 1.7634... gallons)
- The waiting time in line at McDonalds on a given day and time for a given person (can be 2 minutes, 2.3, or 2.34768... minutes)

Discrete distributions

A discrete probability distribution can be described by a table, if it takes finite, values, by a formula, or by a graph.

For example, suppose that *X* is a random variable that represents the number of people waiting at the line at a fast-food restaurant and it happens to only take the values 2, 3, or 5 with probabilities 2/10, 3/10, and 5/10 respectively.

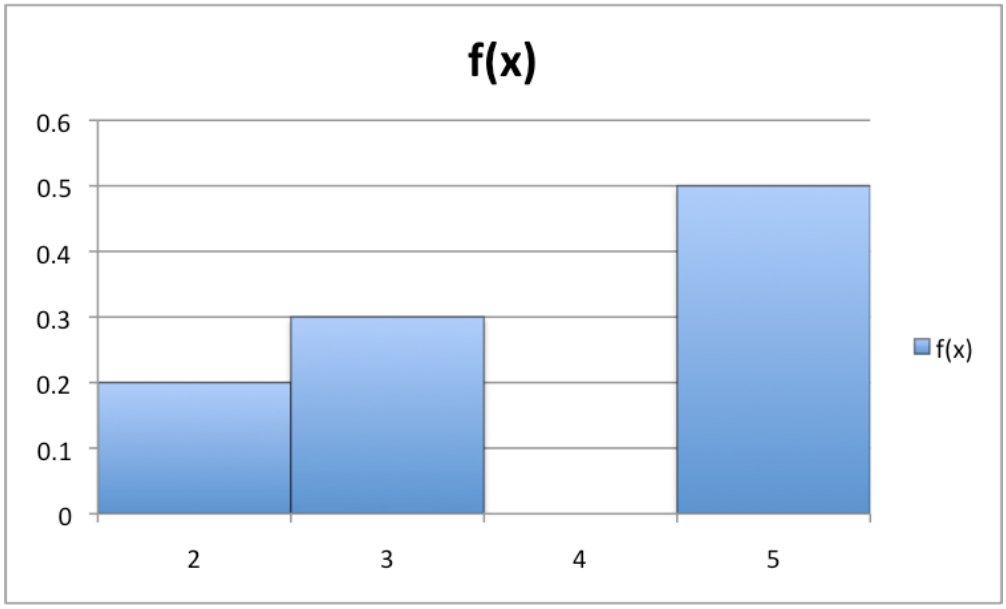
This can be expressed through the function

$$f(x) = x/10, \qquad x = 2, 3, 5 \tag{2}$$

or through the table

x	f(x)
2	0.2
3	0.3
5	0.5

Notice that this two representations are equivalent, and that this can be represented graphically as in the **probability histogram** given below



A probability distribution satisfies the following properties:

- $0 \leq f(x) \leq 1$, i.e., the values of f(x) are probabilities, hence between 0 and 1.
- $\sum f(x) = 1$, i.e., adding the probabilities of all disjoint cases, we obtain the probability of the sample space, 1.

Notice that the cumulative probability distribution function $F(x) = P(X \leq x)$ defined above is related to the probability density function $f(x)$ by

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} f(k). \tag{3}$$

Expected value / Mean

For any probability distribution, one can find the **mean** or **expected value** by calculating

$$\mu_X = \mathbb{E}[X] = \sum xf(x) \tag{4}$$

For the example given in the table above, one could obtain the mean by extending the table as shown below

x	f(x)	xf(x)
2	0.2	0.4
3	0.3	0.9
5	0.5	2.5
sums	1	**3.8**

Therefore,

$$\mu_X = \sum xf(x) = \underline{3.8} \tag{5}$$

Variance

The variance σ_X^2 of a random variable *X* is given by

$$\sigma_X^2 = \mathrm{Var}[X] = \sum (x-\mu)^2 f(x) \tag{6}$$

This can be written equivalently as:

$$\sigma_X^2 = \sum x^2 f(x) - \mu^2 \tag{7}$$

Using the second version of the formula, one can extend the table above to obtain:

x	f(x)	xf(x)	$x^2 f(x)$
2	0.2	0.4	4*0.2 = 0.8
3	0.3	0.9	9*0.3 = 2.7
5	0.5	2.5	25*0.5 = 12.5
sums	1	3.8	16.0

Therefore,

$$\sigma^2 = \sum x^2 f(x) - \mu^2 = 16 - (3.8)^2 = 1.56 \tag{8}$$

And since

> the standard deviation is the square root of the variance,

we obtain

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.56} \approx 1.25 \tag{9}$$

Bernouilli distribution

> A random variable *X* is said to have a **Bernouilli distribution** if *X* takes only two values: X=0 and X=1. "1" represents success (**S**), which has a fixed probability denoted by ***p***, and "0" represents failure (**F**), with fixed probability ***q*=1-*p***. Therefore

$$P(X = 0) = P(F) = q, \tag{10}$$

$$P(X = 1) = P(S) = p \tag{11}$$

and

$$p + q = 1, \quad q = 1 - p, \quad p = 1 - p \tag{12}$$

Binomial distribution

> A random variable *X* is said to have a **Binomial distribution** if it consists of ***n* independent trials**, where each for each trial there is only two options: success (**S**), which has a fixed probability denoted by ***p***, and "0" represents failure (**F**), with fixed probability ***q*=1-*p***. *X* represent the total number of successes, which ranges from 0 to *n*.

The probability of getting exactly *k* successes is given by

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \tag{13}$$

Tables to calculate the binomial distribution are available online (for example here) and in the back of the book.

Example

Q: You answer a multiple choice quiz at random, which has 4 questions, each with 5 options. What is the probability that you get exactly 3 questions correct.

A: For each question, there is only two options: correct or wrong, (success or failure). There is a fixed number of trials *n=4*, the number of questions.

For each question, the probability of success is *p=1/5=0.2*, and the probability of failure is *q=4/5*.

In the problem, we are asked what is the probability of getting exactly 3 answers correct, so *k=3*.

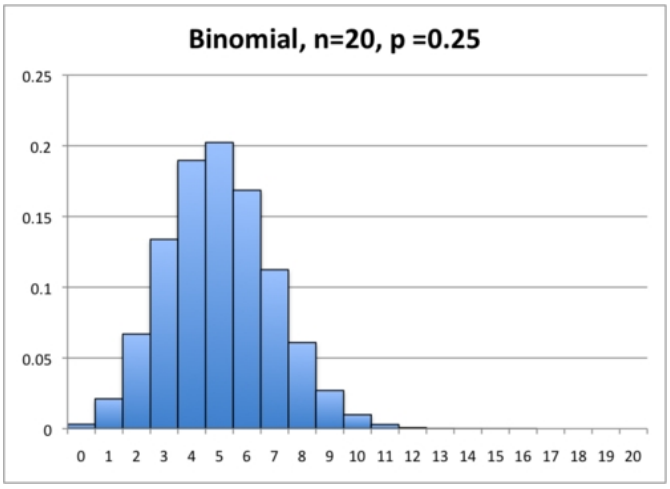
Reading from the table in the link above, we find that for *n=4, k=3*, and *p=0.2*,

$$P(X = k) = 0.0256 \tag{14}$$

We could have found this answer from the definition:

$$P(X=k)=\binom{4}{3}(0.2)^3(0.8)^1=4*0.008*0.8=4*0.0064=0.0256 \quad (15)$$

Below is an example of the probability histogram for a binomial distribution:



Hypergeometric distribution

For the binomial distribution, the independence of the trials was essential.

Suppose that you have an urn with 100 balls, 25 of which are red (success) and 75 are black (failure).

Suppose you sample 4 at random. Notice that here, you would usually be **sampling without replacement**, and hence there is **no independence**. Therefore, the binomial distribution does not apply.

The probability of getting exactly 3 red ones can be obtained as follows:

The number of ways to select 3 red ones out of the 25 red ones is $\binom{25}{3}$=2300.
The number of ways to select 1 black one out of 75 is $\binom{75}{1}$=75.
And in the denominator we have
the number of ways in which you can sample 4 balls out of 100. $\binom{100}{4}$ = 3921225.

Therefore, the probability of getting exactly 3 successes (3 red balls) is

$$P(X=3)=\frac{\binom{25}{3}\binom{75}{1}}{\binom{100}{4}}=0.044 \quad (16)$$

In general, if

> a = number of successes in the population
> b = number of failures in the population, N = a+b
> n = sample size
> x = number of successes in the sample

then

$$P(X=x)=\frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}} \quad (17)$$

> A random variable *X* is said to have a **hypergeometric distribution** if it consists of ***n* dependent trials (with replacement)**, where each for each trial there is only two options: success (**S**). *X* represent the total number of successes, which ranges from 0 to *n*. The number of successes is determined from the number of successes in the population *a*, the number of failures in the population *b*, and the size of the sample *n*.

We mentioned earlier that if the sample is less than 5% of the population, then one can assume independence (replacement) even when the sampling is done without replacement. This leads to the

Binomial approximation of the hypergeometric distribution

> If *n<(0.05) (a+b)*, then one can approximate the hypergeometric distribution with a binomial distribution by setting
> *p=a/(a+b)*
> *q=b/(a+b)*
> n = n

Poisson distribution

The Poisson distribution is the typical jump distribution, where the values (for example people waiting in line at a certain) time, can only jump in discrete amounts from time to time.

It is given by

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, 3, \ldots \tag{18}$$

> It can also approximate the binomial distribution accurately when
> n>100
>
> np<10

Then $\lambda = np$, so we obtain:

$$f(x) = e^{-np}\frac{(np)^x}{x!}, \quad x = 0, 1, 2, 3, \ldots \tag{19}$$

Example

The number of people on the registrar's office line at a college 3 days after registration follows a Poisson distribution with parameter λ=5.

1. Find the probability that a person finds 4 people in line when arriving at the registration office.
2. Find the probability that a person finds at least 4 people in line when arriving at the registration office.

1. Let X be the random variable denoting the number of people waiting in line. Then

$$P(X = 4) = e^{-\lambda}\frac{\lambda^4}{4!} = e^{-5}\frac{5^4}{4!} = 0.175 \tag{20}$$

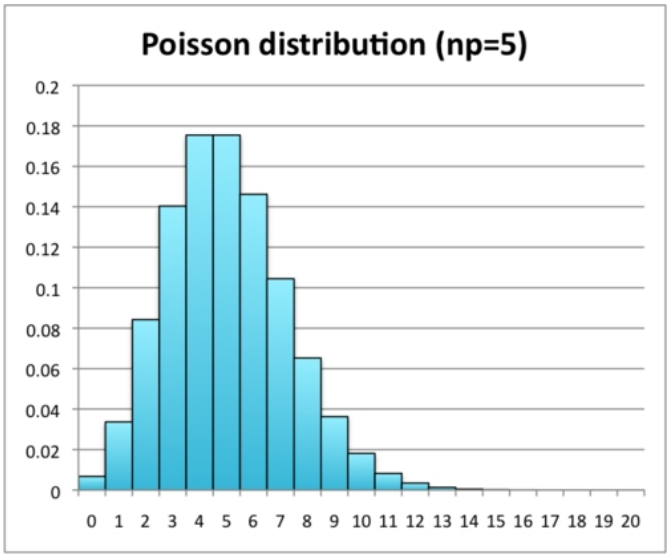
2. The probability of finding at least 4 people in line is the probability of finding 4, 5, 6, 7, 8, 9, 10, … , or more people in line. Since the Poisson distribution can take any non-negative integer value, this probability would take too long to calculate by finding each one of these individual probabilities. Instead, it would be easier to find the probability of the complement:

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) + P(X = 6) + \ldots \\ &= 1 - P(X < 4) \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)] \\ &= 1 - [0.007 + 0.034 + 0.084 + 0.140] = 0.735 \end{aligned} \tag{21}$$

The values in the following table were calculated using the excel formula POISSON((x,mean, cummulative):

x	0	1	2	3	4	5	6	7	8	...
f(x)	0.007	0.034	0.084	0.14	0.175	0.175	0.15	0.10	0.07	...

Below is the probability histogram for the Poisson distribution in this example:



Notice that for the parameters given, this probability distribution is similar to the one for the binomial distribution example, though not the same. Remember that the Poisson distribution is good at approximating

the binomial distribution when n>100 and np<10.