Probability Distribution of A Random Variable (*r.v.*)

Huining Kang

HuKang@salud.unm.edu

July 27, 2020

Contents

- Probability distributions of discrete variables
  - Basic concepts of probability distributions
    - Probability distribution of a random variable
    - Probability density function (pdf)
    - Cumulative probability function (cdf)
    - Characteristics of a discrete random variable
  - Probability calculation based on pdf or cdf

Objective

- Understand the basic concepts of probability distributions of discrete variables
- Be able to calculate probabilities of a discrete random variable based on the pdf or cdf.

Probability distributions of discrete variables – Basic Concepts

- Concerns about the random experiments
  - Possible Outcomes (can be expressed in the form of random variable)
  - Probabilities of the outcomes
- Probability distribution
  - The relationship between the outcomes (value of a random variable) and the probability of the their occurrence is referred to as probability distribution.
  - Probability distribution may be expressed in the form of a table, graph or formula
- Two important concepts
  - Probability density function (pdf)
  - Cumulative distribution function (cdf)

Probability distributions of discrete variables – Basic Concepts

- Example 1 (p93 4th Ed., example 4.2.1)

Data	
Family ID	# of programs
1	3
2	1
3	8
:	:
:	:
296	6
297	*4*

EXAMPLE 4.2.1

- Source: Holben et al. (2004) Journal of the American Dietetic Association
- Investigation: food security status in families in the Appalachian region of southern Ohio.
- Study purpose: to examine hunger rates of families with children in a local Head Start program
- The survey instrument included the 18-questions regarding hunger and food security.
- An addition questions: participants were asked how many food assistance programs they had used in the last 12 months.

# of Programs	1	2	3	4	5	6	7	8
Frequency	62	47	39	39	58	37	4	11

- Experiment: pick one family at random, look at the # of programs they used.

Probability distributions of discrete variables – Basic Concepts

- Example 1 (Cont.)

Data	
Family ID	# of programs
1	3
2	1
3	8
:	:
:	:
296	6
297	*4*

- Experiment: pick one family at random
- Outcome of Interest: Number of assistance programs utilized by the family(X)

  The values of X : 1, 2, 3, 4, 5, 6, 7, 8 (x)
- Probability density function (pdf):

$$p(x) = \mathrm{P}\,(X = x),\ \ x = 1, 2, \ldots, 8$$

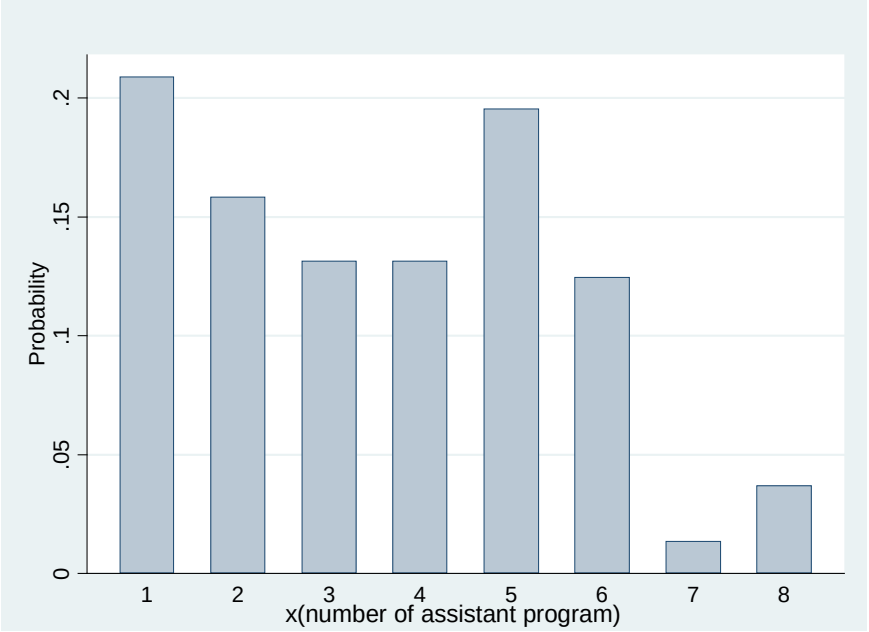
  Properties: (1) $p(x) \geq 0$, (2) $\Sigma_x\, p(x) = 1$
- Cumulative distribution function (cdf):

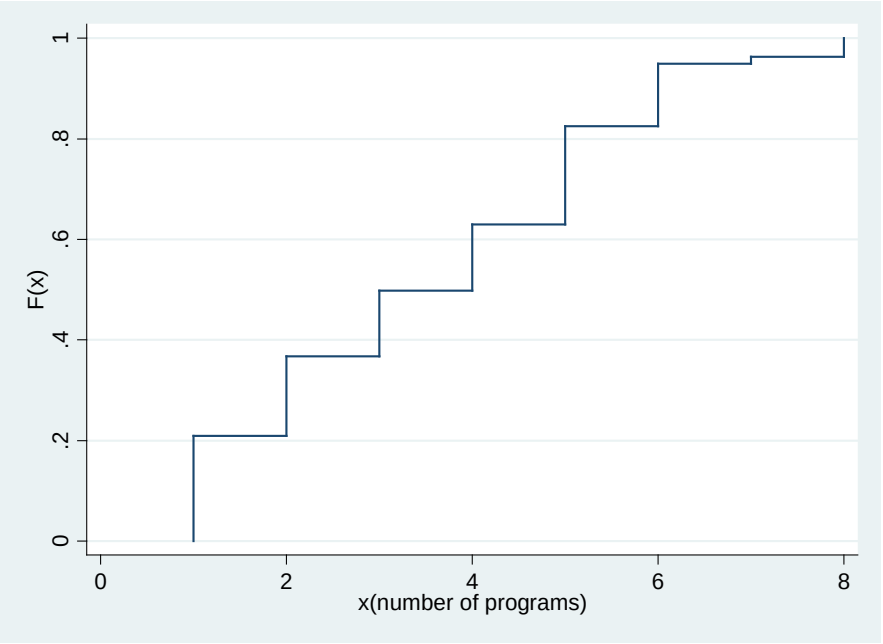
$$F(x) = \mathrm{P}\,(X \leq x), \text{ for all } x$$

- $$p(x) = F(x) - F(x-1)$$

X	1	2	3	4	5	6	7	8	x
Freq	62	47	39	39	58	37	4	11	
Frac	.2088	.1582	.1313	.1313	.1953	.1246	.0135	.0370	$p(x)$
Cum	.2088	.3670	.4983	.6296	.8249	.9495	.9630	1.000	$F(x)$

Probability distributions of discrete variables – Basic Concepts





Probability density function

$p(x) = P(X = x)$

Cumulative distribution function

$F(x) = P(X \leq x)$

X	1	2	3	4	5	6	7	8	x
Freq	62	47	39	39	58	37	4	11	
Frac	.2088	.1582	.1313	.1313	.1953	.1246	.0135	.0370	$p(x)$
Cum	.2088	.3670	.4983	.6296	.8249	.9495	.9630	1.000	$F(x)$

Probability distributions of discrete variables – Basic Concepts

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	

Average number of programs used:

$$\mu = \frac{(1\times62+2\times47+\cdots+8\times11)}{297} = 1 \times .2088 + 2 \times .1582 + \cdots + 8 \times .0370 = 3.56$$

Variance: $\sigma^2 = (1 - \mu)^2 \times .2088 + (2 - \mu)^2 \times .1582 + \ldots + (8 - \mu)^2 \times .0370 = 3.86$,
Standard deviation (SD): $\sigma = 1.96$

μ, σ^2 and σ are called mean (or expectation), variance and standard deviation of X

Probability distributions of discrete variables – Basic Concepts

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	

Average number of programs used:

$$\mu = 1 \times .2088 + 2 \times .1582 + \ldots + 8 \times .0370 = 3.56 \ (\mathrm{E}(X))$$

Variance: $\sigma^2 = (1 - \mu)^2 \times .2088 + (2 - \mu)^2 \times .1582 + \ldots + (8 - \mu)^2 \times .0370 = 3.86$ (Var(X)),
Standard deviation (SD): $\sigma = 1.96$

μ, σ^2 and σ are called mean (or expectation), variance and standard deviation of X

Probability distributions of discrete variables – Basic Concepts

Distribution of Random Variable X

X	x_1	x_2	…	x_k	…	x_n
P	p_1	p_2	…	p_k	…	p_n

The mean (mathematical expectation, or expected value) of X :

$$\mu = E(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_k \cdot p_k + \ldots + x_n \cdot p_n$$

or

$$\mu = E(X) = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \ldots + x_k \cdot p(x_k) + \ldots + x_n \cdot p(x_n)$$

Variance of X :

$$\sigma^2 = Var(X) = (x_1 - \mu)^2 \times p_1 + (x_2 - \mu)^2 \times p_2 + \ldots + (x_k - \mu)^2 \times p_k + \ldots + (x_n - \mu)^2 \times p_n$$

Standard deviation of X :

$$\sigma = \sqrt{Var(X)}$$

Probability distributions of discrete variables – Calculation of the probability

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	

EXAMPLE 4.2.2

What is the probability that a randomly selected family used three assistance programs?

SOLUTION:

$$p(3) = P(X = 3) = .1313$$

Probability distributions of discrete variables – Calculation of the probability

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	

EXAMPLE 4.2.3

What is the probability that a randomly selected family used either one or two programs?

SOLUTION:

$$P("X = 1" \cup "X = 2") = P("X = 1") + P("X = 2") = p(1) + p(2)$$

$$= .2088 + .1582 = .3670$$

Probability distributions of discrete variables – Calculation of the probability

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	

EXAMPLE 4.2.4

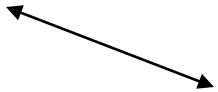
What is the probability that a family picked at random used two or fewer assistance programs?

SOLUTION:

$$P(X \leq 2) = F(2) = \quad .3670$$

Probability distributions of discrete variables – Calculation of the probability

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	



EXAMPLE 4.2.6

What is the probability that a randomly selected family used five or more programs?

SOLUTION:

$$P(X \geq 5) = 1 - F(X \leq 4) = 1 - .6296 = .3704$$

Probability distributions of discrete variables – Calculation of the probability

		Relative frequency	Cumulative frequency
Number of programs (x)	Frequency	$p(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	62	.2088	.2088
2	47	.1582	.3670
3	39	.1313	.4983
4	39	.1313	.6296
5	58	.1953	.8249
6	37	.1246	.9495
7	4	.0135	.9630
8	11	.0370	1.000
Total	297	1.000	

Probability that the number of programs used is between 3 and 5 inclusive:

$$P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5) = p(3) + p(4) + p(5) = 0.4579$$

$$P(2 < X \leq 5) = P(X \leq 5) - P(X \leq 2) = F(5) - F(2) = 0.4579$$

Probability distributions of discrete variables – Calculating Probabilities Using Stata

```
* Example 4.2.1

* Imort data set to Stata memory
import delimited using "data/EXA_C04_S02_01.csv", clear

* Take a look at data
describe
list

* Rename the variable to a shorter name
rename numberofprograms nprg

* Take a look at data, again
describe
list

* Calculate relative percentage and cummulative percentage
tabulate nprg [fweight = frequency]

* Plot probability density function
histogram nprg [fweight = frequency], discrete color(blue)
fintensity(inten10) barwidth(0.8) xlabel(#8)
```

Probability distributions of discrete variables – Calculating Probabilities Using Stata

```
* Generate new variables for relative frequency
* i.e., Probability density function

egen totalFreq = sum(frequency)
gen Frac = frequency/totalFreq
label variable Frac "Relative Frequency"
drop totalFreq

* Take a look at data, again
list

* Generate cumulative distribution
cumul nprg [fweight = frequency], generate(cum)
* Take a look at data, again
List

* Calculate mean, variance and standard deviation
summarize nprg [fweight = frequency], detail
```

Probability distributions of discrete variables – Calculating Probabilities Using Stata

```
* Make the plot of cumulative distribution

gen cum1 = cum[_n - 1]
replace cum1=0 if cum1==.
rename cum cum2
reshape long cum, i(nprg) j(idx)
label variable cum "Probability"
set obs 17
replace nprg = 0 if nprg==.
replace cum = 0 if cum==.
set obs 18
replace nprg = 9 if nprg==.
replace cum = 1 if cum==.
sort nprg idx
line cum nprg, sort xlabel(#10)
```

Summary

- Contents
  - Basic concepts of probability distributions
    - Probability distribution of a random variable
    - Probability density function (pdf)
    - Cumulative probability function (cdf)
    - Characteristics of a discrete random variable
  - Probability calculation based on pdf or cdf
- Objective
  - Understand the concepts
  - Know how to calculate the probability based on cdf and pdf

Homework 4

- Due August 10, 2020
- Exercises 4.2.1, 4.2.2, 4.2.3 (Page 98, 4th Ed).
- Data file: EXR_C04_S02_01
- Hint: Modify the Stata do file lecture05.do to complete the exercise 4.2.1 and 4.2.3. Do the calculation for 4.2.2 by hand.