Lecture - Day 3 (part 1)
Binomial distribution

MF9130E V24
2024.04.10

Chi Zhang
Oslo Center for Biostatistics and Epidemiology
chi.zhang@medisin.uio.no

Outline

Discrete and continuous variables, probability mass and density

Probability distributions

Binomial distribution

Normal distribution, standard normal distribution, quantiles

CLT: central limit theorm, useful for inference

Random variables

Discrete and continuous

Random variable: a quantity that can take random values, with a certain probability

Discrete variables
- coin tossing H,T
- birth boy, girl

Continuous variables
- weight and height
- age

Properties of probability:

Non-negative (0 or above), less than 1, prob of all outcomes sum up to 1.

Histogram and bar plot
X-axis is usually what **values** the variable can be - either a single value, or a range

Y-axis is the **frequency**; or **proportions** (probability) corresponding to that variable value (or range)

Bar plot

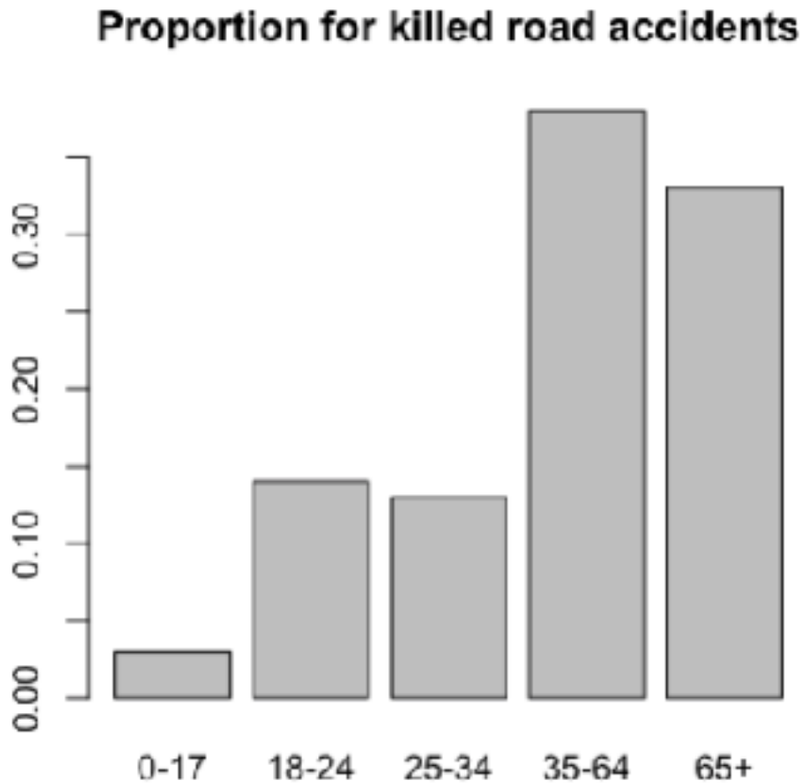
Road accident example

Probability distribution: relationship between outcome and probability

Make a visualization for the road accident data, by **age group**: only y-axis changed

Proportions will sum up to 1 (up to a rounding error)

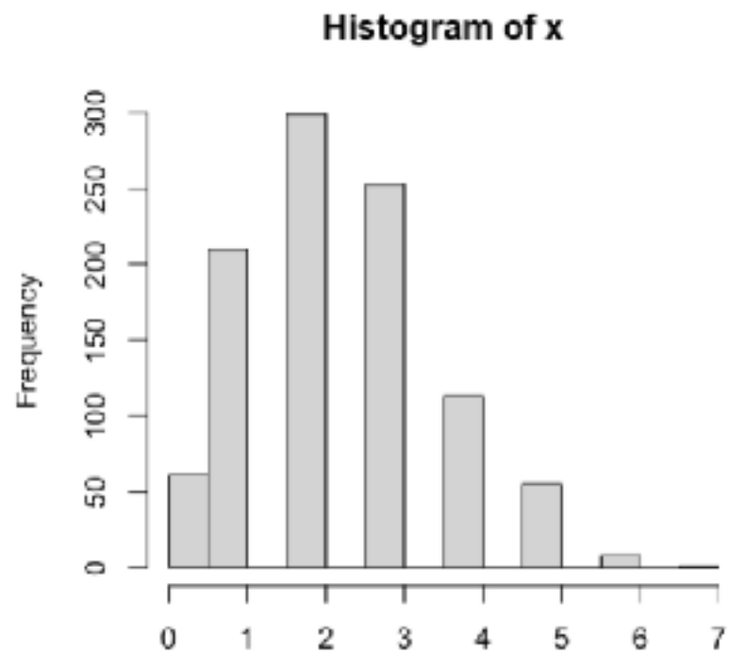
Age group	Number of death	Proportion among all deaths
0-17	3	3/116 = 0.03
18-24	16	0.14
25-34	15	0.13
35-64	44	0.38
65 and above	38	0.33
Total	116	



Probability distribution

Discrete variables

You use **histograms** or **bar plots** to have an impression of where your data lies, with what probability



With theoretical models, you can compute the exact probability of a random variable X takes x values

For **discrete** variable X, it only has probability at certain values (0,1,2, …, 7).
We can compute P(X=0), P(X=1), …
Either from formula (if you know the theoretical distribution), or divide the counts by N

Can also find X equals to a range of values by **aggregating groups**:
P(X>5) = P(X=6) + P(X=7)

For **continuous** variables, X can take any value: 0, 0.5, 0.55, 0.5555… can also be negative. How do we find the probabilities? -> next lecture.

Binomial distribution

On day 4 we introduce counts and proportions

A proportion can be seen as the probability of **getting exactly d events in a sample of n**.

For example, toss a coin 100 times, assuming that its fair (equal probability for Heads and Tails - proportion is 0.5), then we can expect to get 50 Heads.

Independent experiments: each toss does not affect another
P(X = H) is the same for all experiments

Two outcomes (1, 0; positive, negative; Heads Tails)
P(X = H) + P(X = T) = 1 (complementary rule)

Binomial distribution

Now we are interested in the number of H when we toss a coin for once, twice, three times. We denote Heads as 1, and Tails as 0

What are the possible outcomes for these four situations?
Assume P(1) = 0.5, P(0) = 0.5

One toss situation
2 results, either 1 or 0

Toss 1	Probability
1	0.5
0	0.5

Toss twice situation
4 results (2^2)

Both 1, first 1 second 0, first 0 second 1, both 0

Toss 1	Toss 2	Probability
1	1	0.5*0.5 = 0.25
1	0	0.5*0.5 = 0.25
0	1	0.5*0.5 = 0.25
0	0	0.5*0.5 = 0.25

Binomial distribution

Toss three times:
8 results (2^3)

Toss 1	Toss 2	Toss 3	Probability
1	1	1	0.5*0.5*0.5 = 0.125
1	1	0	0.125
1	0	1	0.125
1	0	0	0.125
0	1	1	0.125
0	1	0	0.125
0	0	1	0.125
0	0	0	0.125

Binomial distribution

What are the probability of getting **exactly zero 1, one 1, two 1, three 1**?

Toss 1	Probability
1	0.5
0	0.5

P(zero 1) = 0.5
P(one 1) = 0.5

Toss 1	Toss 2	Probability
1	1	0.5*0.5 = 0.25
1	0	0.5*0.5 = 0.25
0	1	0.5*0.5 = 0.25
0	0	0.5*0.5 = 0.25

P(zero 1) = 0.25

P(one 1) = 0.25 + 0.25

P(two 1) = 0.25

Binomial distribution

What are the probability of getting exactly **zero 1, one 1, two 1, three 1**?

Toss 1	Toss 2	Toss 3	Probability
1	1	1	0.5*0.5*0.5 = 0.125
1	1	0	0.125
1	0	1	0.125
1	0	0	0.125
0	1	1	0.125
0	1	0	0.125
0	0	1	0.125
0	0	0	0.125

P(zero 1) = 0.125

P(one 1) = 0.125*3 = 0.375

P(two 1) = 0.125*3 = 0.375

P(three 1) = 0.125

In n trials (experiments), the probability of event X occurs exactly x times is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

binomial coefficient,C(n,x)

Binomial distribution

8 patients come to do screening tests for disease A.

Assume these patients are unrelated with each other. Their probability of getting a positive result is 0.15.

What is the probability of 2 patients have positive results for disease A?
What about probability of 5 patients?

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

For n = 8 (total number of patients)

P(X = 2, disease A) = C(8, 2) * 0.15^2 * (1-0.15) ^ (8-2) = 0.238
P(X = 5, disease A) = C(8, 5) * 0.15^5 * (1-0.15) ^ (8-5) = 0.0026

Theoretical or empirical?

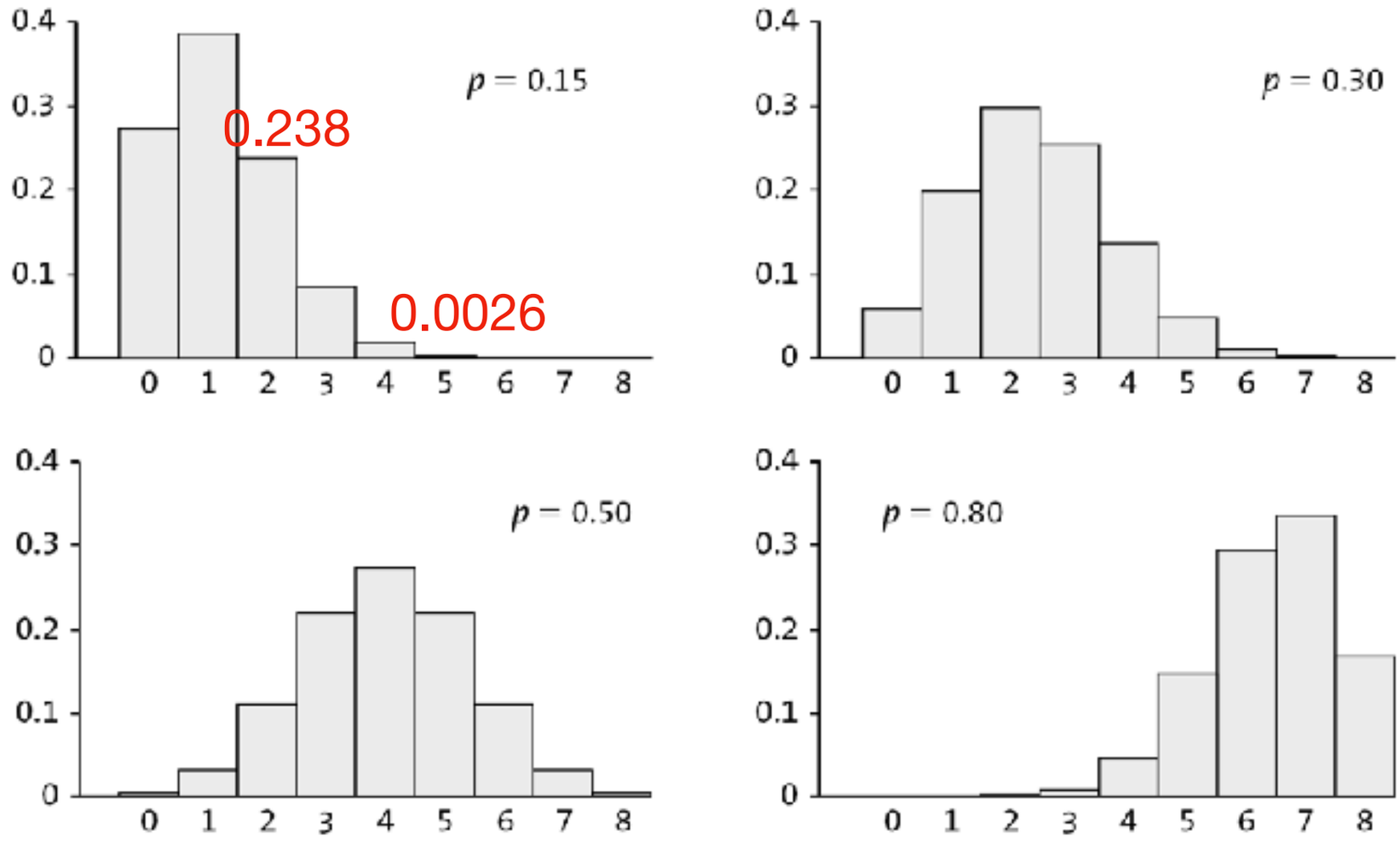
Binomial distribution is a **theoretical** distribution -> we compute the probability given certain parameters

e.g. n and p for binomial distribution

When we collect data, divide data into groups, make plots, we are making an **empirical** approximation. There is randomness, and the approximation is not exactly the same as the theoretical distribution.

The approximation can be quite useful, as you will see in the next lecture on normal distribution.

Binomial distribution



Figur 4.2 Histogrammer for binomisk fordeling med $n = 8$ forsøk og fire forskjellige verdier av p

Binomial distribution can be approximated by **normal distribution** under some conditions (if np>10, n(1-p)>10, p close to 0.5). The approximation is **N(np, np(1-p))**

This result is relevant for making statistical inference (uncertainty of proportions, z-test)