Lecture 05: Probability Distributions

Applied Statistics - PKN STAN - 537 - 538

Lecturer: Erika Siregar, SST, MS

27 & 31 October 2020

Review

1. Apa itu probability
2. Apa saja 4 cara menentukan sampel space?
3. Apa itu kombinasi → C
4. Apa itu permutasi
5. Apa itu peluang bersyarat → $P(X|Y)$
6. Apa ciri dari kejadian independent → $P(A \text{ and } B) = P(A)*P(B)$
7. Apa ciri dari kejadian mutually exclusive. →

Data Distribution???

Probability Distribution???

- Diketahui peluang per kejadian.
- How about the whole events?

The Background Story (2)

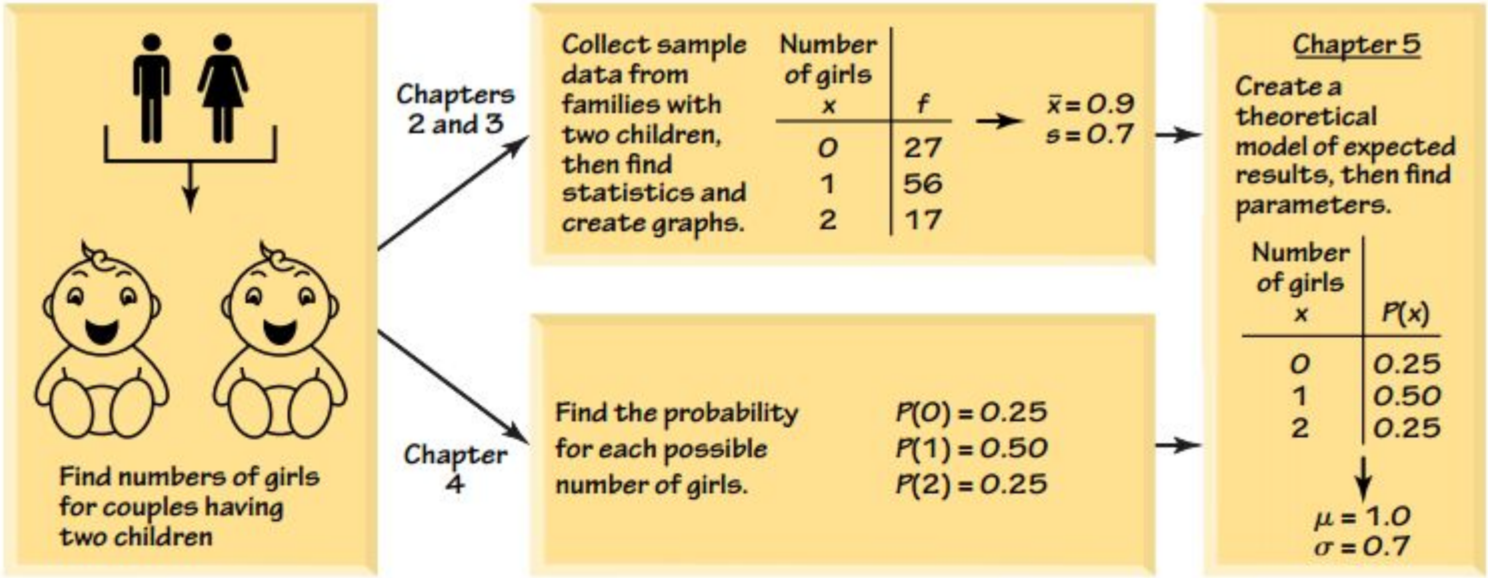
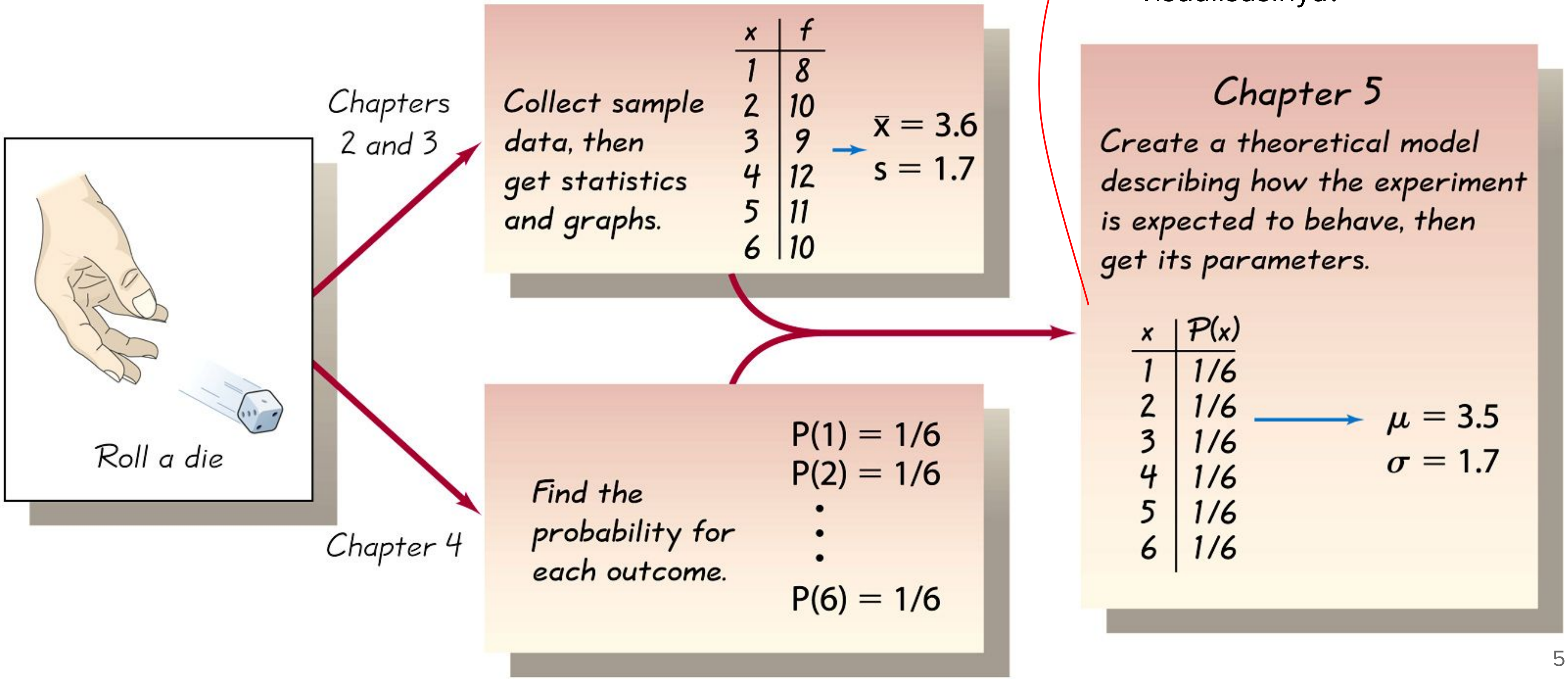


Figure 5-2

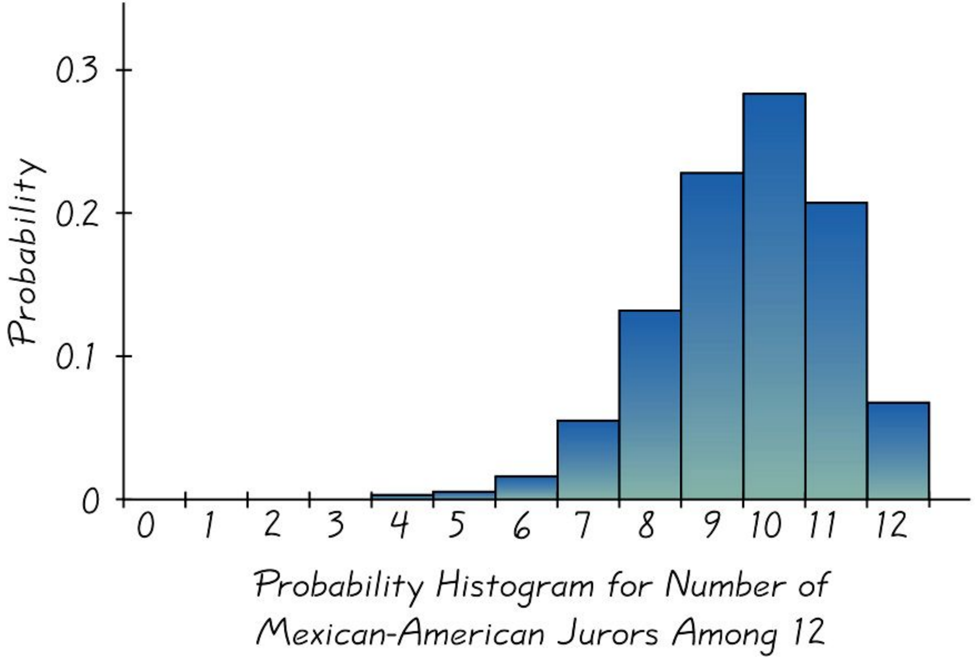
The Background Story (2)

Tabel frekuensi. Bgmn visualisasinya?



Probability Histogram

similar to a relative frequency histogram, but the **vertical scale (sumbu y)** shows **probabilities**.



Random Variable

- Event yang akan kita teliti probability-nya. ➔ represents all possibilities in the entire sample space
- Misal: jumlah bayi perempuan yang lahir, jumlah mata dadu yang keluar, dll
- **Harus numerik**
- Further distinguished into: discrete & continuous.
- Probability distribution ➔ random variable

Probability Distributions

- sebaran probability across all possibilities in an event (random variable).
- describe **what will probably happen** at any particular number of events instead of what actually did happen.
- often given in the format of a graph, table, or formula.
- untuk kasus tertentu bisa di-generalize ke dalam formula.

Probability Distributions Requirements

1. There is a **numerical random variable x** and its values are **associated with corresponding probabilities**.
2. $\Sigma P(x_i) = 1$ → (0.999 or 1.001 are acceptable because they result from rounding errors.)
3. $0 <= P(x_i) <= 1$

Table 5-3 When to Discuss Salary

Number of Interviews x	P(x)
1	0.30
2	0.26
3	0.10

x	P(x)
1	0.001
2	0.020
3	0.105
4	0.233
5	0.242

Table 5-2 Should Marijuana Use Be Legal?

Response	P(x)
Yes	0.41
No	0.52
Don't know	0.07

Table 5-1 Probability Distribution for the Number of Girls in Two Births

Number of Girls x	P(x)
0	0.25
1	0.50
2	0.25

Example

Example 1 Genetics

Although the Chapter Problem involves 945 births, let's consider a simpler example that involves only two births with the following random variable:

$$x = \text{number of girls in two births}$$

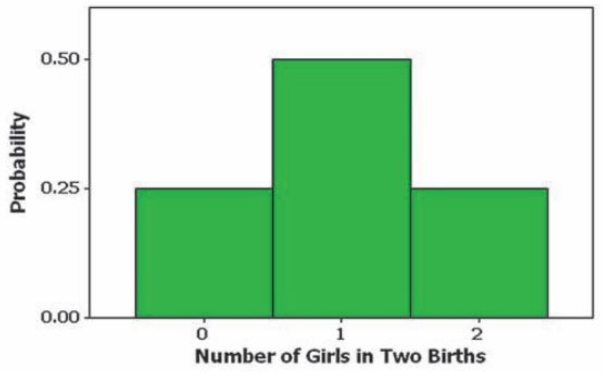
The above x is a random variable because its numerical values depend on chance.

Table 5-1 Probability Distribution for the Number of Girls in Two Births

Number of Girls x	$P(x)$
0	0.25
1	0.50
2	0.25

It satisfies the 3 requirements for a probability distribution

1. The variable x is a numerical random variable and its values are associated with probabilities, as in Table 5-1.
2. $\Sigma P(x) = 0.25 + 0.50 + 0.25 = 1$
3. Each value of $P(x)$ is between 0 and 1. (Specifically, 0.25 and 0.50 and 0.25 are each between 0 and 1 inclusive.)



Parameters of Prob Distributions

1. **Mean (μ) or Expected Value (E)**
   a. Nilai harapan = kejadian yang paling mungkin muncul secara rata-rata.
   b.

$$\mu = \frac{\Sigma(f \cdot x)}{N} = \sum\left[\frac{f \cdot x}{N}\right] = \sum\left[x \cdot \frac{f}{N}\right] = \Sigma\left[x \cdot P(x)\right]$$

$$\boxed{\mu = \Sigma\left[x \cdot P(x)\right]}$$

2. Variance

$$\sigma^2 = \Sigma\left[(x-\mu)^2 \cdot P(x)\right] \Rightarrow \sigma^2 = \Sigma\left[x^2 \cdot P(x)\right] - \mu^2 \Rightarrow E(X^2) - \mu^2$$

3. Standard Deviation → square root of variance.

$$\sigma = \Sigma\left[x^2 \cdot P(x)\right] - \mu^2$$

Example of Usage

$$\sigma = \Sigma\,[(x - \mu)^2 \cdot P(x)]$$

Table 5-4 Calculating μ and σ for a Probability Distribution

x	***P*(x)**	**x · *P*(x)**	**(x − μ)² · *P*(x)**
0	0.25	$0 \cdot 0.25 = 0.00$	$(0 - 1)^2 \cdot 0.25 = 0.25$
1	0.50	$1 \cdot 0.50 = 0.50$	$(1 - 1)^2 \cdot 0.50 = 0.00$
2	0.25	$2 \cdot 0.25 = 0.50$	$(2 - 1)^2 \cdot 0.25 = 0.25$
Total		1.00 ↑ $\mu = \Sigma[x \cdot P(x)]$	0.50 ↑ $\sigma^2 = \Sigma[(x - \mu)^2 \cdot P(x)]$

Apa interpretasinya?

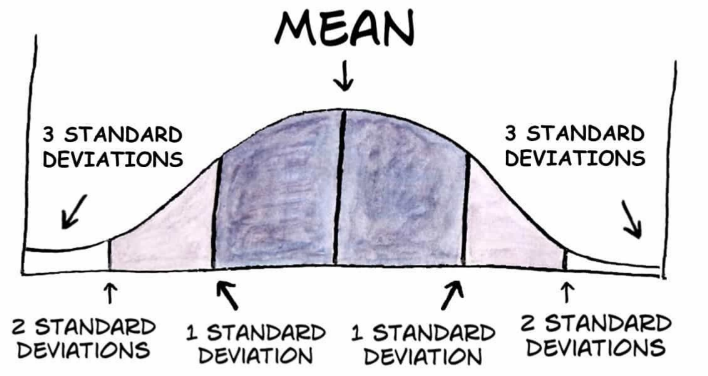
Identifying Unusual

I. Using μ and σ

Range Rule of Thumb

$$\text{maximum usual value} = \mu + 2\sigma$$

$$\text{minimum usual value} = \mu - 2\sigma$$



II. using probability

- Unusually high: x successes among n trials is an unusually high number of successes if **P(x or more) ≤ 0.05**. ➔ P(>= X)
- Unusually low: x successes among n trials is an unusually low number of successes if **P(x or fewer) ≤ 0.05**. P(<= x)

Contoh:

Banyaknya mobil yang masuk ke tempat cuci mobil setiap hari antara jam 1 pm - 2 pm memiliki distribusi peluang sebagai berikut. Berapakah expected value dan variance dari disribusi peluang ini?

x	P(x)	x.P(x)	(x-μ)^2*P(x)
0	0.2	0	0.8
1	0.1	0.1	0.1
2	0.3	0.6	0
3	0.3	0.9	0.3
4	0.1	0.4	0.4
		2	1.6

$$\boxed{\mu = \Sigma\left[x \cdot P(x)\right]}$$

$$\sigma^2 = \Sigma\left[(x - \mu)^2 \cdot P(x)\right]$$

Try it with R ➔ weighted.mean()

Contoh 2

Genetics. ***In Exercises 15–18, refer to the accompanying table, which describes results from groups of 10 births from 10 different sets of parents. The random variable x represents the number of girls among 10 children.***

Number of Girls x	P(x)
0	0.001
1	0.010
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.010
10	0.001

15. Mean and Standard Deviation Find the mean and standard deviation for the numbers of girls in 10 births.

16. Range Rule of Thumb for Unusual Events Use the range rule of thumb to identify a range of values containing the usual numbers of girls in 10 births. Based on the result, is 1 girl in 10 births an unusually low number of girls? Explain.

17. Using Probabilities for Unusual Events

a. Find the probability of getting exactly 8 girls in 10 births.

b. Find the probability of getting 8 or more girls in 10 births.

c. Which probability is relevant for determining whether 8 is an unusually high number of girls in 10 births: the result from part (a) or part (b)?

d. Is 8 an unusually high number of girls in 10 births? Why or why not?

18. Using Probabilities for Unusual Events

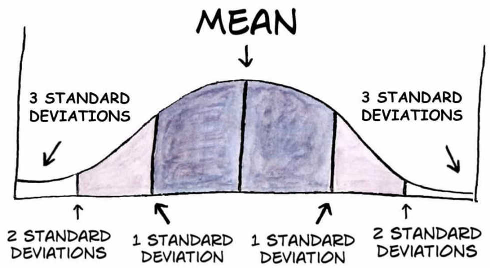
a. Find the probability of getting exactly 1 girl in 10 births.

b. Find the probability of getting 1 or fewer girls in 10 births.

c. Which probability is relevant for determining whether 1 is an unusually low number of girls in 10 births: the result from part (a) or part (b)?

d. Is 1 an unusually low number of girls in 10 births? Why or why not?

Contoh 2 (solution)



x	P(x)	x.P(x)	(x-μ)^2*P(x)	x^2	x^2 * P(x)
0	0.001	0	0.025	0	0
1	0.01	0.01	0.16	1	0.01
2	0.044	0.088	0.396	4	0.176
3	0.117	0.351	0.468	9	1.053
4	0.205	0.82	0.205	16	3.28
5	0.246	1.23	0	25	6.15
6	0.205	1.23	0.205	36	7.38
7	0.117	0.819	0.468	49	5.733
8	0.044	0.352	0.396	64	2.816
9	0.01	0.09	0.16	81	0.81
10	0.001	0.01	0.025	100	0.1
sum	1	5	2.508	385	27.508

- Is 1 girl in 10 births unusual?
  - (5-2.508) <= x <= (5+2.508)
  - 2.492 <= x <= 7.508

- 17a = 0.044
- 17b = 0.044 + 0.01 + 0.001 = 0.055
- **Unusually high jika P(x or more)** ≤ 0.05

➜ P(x >= 8) = 0.055??

- 18a = 0.01
- 18b = 0.001 + 0.01 = 0.011
- **Unusually low jika P(x or fewer) ≤ 0.05.**

➜ P(x <= 1) = 0.011 ??

Try it with R

Special Probability Distributions

- Some probability distributions can be generalized into formulas
- List of prob dists: https://en.wikipedia.org/wiki/List_of_probability_distributions
- For now:
  - Binomial
  - Multinomial
  - Poisson

Binomial Probability Distribution

- Karakteristik kasus: **binary** ➔ 0/1, yes/no, sukses/gagal, dll.
- describes the number of successes in a series of independent Yes/No experiments all with the same probability of success.
- Requirement:
    - The procedure has a fixed number of trials
    - The trials must be independent
    - outcomes classified into two categories for outcomes ➔ **yes/no, success/failure, 0/1**
    - The probability of a success remains the same in all trials

Example 1 Twitter

When an adult is randomly selected (with replacement), there is a 0.85 probability that this person knows what Twitter is (based on results from a Pew Research Center survey). Suppose that we want to find the probability that exactly three of five random adults know what Twitter is.

a. Does this procedure result in a binomial distribution?

b. If this procedure does result in a binomial distribution, identify the values of n, x, p, and q.

Solution

a. This procedure does satisfy the requirements for a binomial distribution, as shown below.

1. The number of trials (5) is fixed.
2. The 5 trials are independent, because the probability of any adult knowing Twitter is not affected by results from other selected adults.
3. Each of the 5 trials has two categories of outcomes: The selected person knows what Twitter is or that person does not know what Twitter is.
4. For each randomly selected adult, there is a 0.85 probability that this person knows what Twitter is, and that probability remains the same for each of the five selected people.

b. Having concluded that the given procedure does result in a binomial distribution, we now proceed to identify the values of n, x, p, and q.

1. With five randomly selected adults, we have $n = 5$.
2. We want the probability of exactly three who know what Twitter is, so $x = 3$.
3. The probability of success (getting a person who knows what Twitter is) for one selection is 0.85, so $p = 0.85$.
4. The probability of failure (not getting someone who knows what Twitter is) is 0.15, so $q = 0.15$.

Binomial Distribution Formula

The number of outcomes with exactly x successes among n trials

The probability of x successes among n trials for any one particular order

Formula 5-5 Binomial Probability Formula

$$P(x) = \underbrace{\frac{n!}{(n-x)!x!}}_{\text{The number of outcomes with exactly x successes among n trials}} \cdot \underbrace{p^x \cdot q^{n-x}}_{\text{The probability of x successes among n trials for any one particular order}} \qquad \text{for } x = 0, 1, 2, \ldots, n$$

where

n = number of trials 7

x = number of successes among n trials 4

p = probability of success in any one trial 0.8

q = probability of failure in any one trial ($q = 1 - p$) 0.2

$\boxed{{}^nC_x * p^x * q^{n-x}}$

Appendix A of Triola's book provides the value for some binomial probabilities.

Binomial Distribution

Mean $\mu = n \cdot p$

Variance $\sigma^2 = n \cdot p \cdot q$

Std. Dev. $\sigma = \sqrt{n \cdot p \cdot q}$

$$\boxed{{}^{n}C_{x} * p^{x} * q^{n-x}}$$

5 C 3 * (0.6)^3 * (0.4)^2

Example 4 **Devil of a Problem**

Based on a recent Harris poll, 60% of adults believe in the devil. Assuming that we randomly select five adults, use Table A-1 to find the following:

a. The probability that exactly three of the five adults believe in the devil

b. The probability that the number of adults who believe in the devil is at least two

Solution

a. The following excerpt from the table shows that when $n = 5$ and $p = 0.6$, the probability for $x = 3$ is given by $P(3) = 0.346$.

TABLE A-1 Binomial Probabilities

n	x	.01	...	.50	.60	.70
					p	
5	0	.951		.031	.010	.002
	1	.048		.156	.077	.028
	2	.001		.312	.230	.132
	3	0+		.312	.346	.309
	4	0+		.156	.259	.360
	5	0+		.031	.078	.168

x	P(x)
0	0.010
1	0.077
2	0.230
3	0.346
4	0.259
5	0.078

b. The phrase "at least two" successes means that the number of successes is 2 or 3 or 4 or 5.

$$\begin{aligned} P(\text{at least 2 believe in the devil}) &= P(2 \text{ or } 3 \text{ or } 4 \text{ or } 5) \\ &= P(2) + P(3) + P(4) + P(5) \\ &= 0.230 + 0.346 + 0.259 + 0.078 \\ &= 0.913 \end{aligned}$$

Try it with R

Multinomial Distribution

- Mirip binomial, tapi setiap event dapat menghasilkan > 2 kemungkinan.
- Formula:

$$\frac{n!}{x_1!\,x_2!\,x_3!\cdots x_k!}\,p_1^{x_1}\,p_2^{x_2}\cdots p_k^{x_k}$$

Contoh Multinomial

$$\frac{n!}{x_1!\,x_2!\,x_3!\ldots x_k!}\cdot p_1^{x_1}\,p_2^{x_2}\ldots p_k^{x_k}$$

Bila dua dadu dilantunkan 6 kali, berapakah peluang mendapat jumlah 7 atau 11 muncul **dua** kali, sepasang bilangan yang sama **satu** kali, dan kombinasi lainnya 3 kali?

Answer:

X1 = jumlah 7 atau 11 yang muncul ➔ p = 48/**216** = 2/9

X2 = pasangan bilangan yang sama muncul ➔ 36/**216** ➔ ⅙ ➔ (1,1), ...(6,6)

X3 = kombinasi lainnya ➔ 132/**216** = 11/18

$$\frac{6!}{2!\,1!\,3!}\cdot\left(\frac{2}{9}\right)^2\cdot\left(\frac{1}{6}\right)^1\cdot\left(\frac{11}{18}\right)^3$$

$$= 0{,}1127$$

(1,6), (2,5), (3,4), (4,3), (6,1), (5,2), (5,6), (6,5) ➔ 8*6

Try it with R

Poisson Distribution

- Characteristics: banyaknya sukses yang terjadi dalam **selang waktu** tertentu.
    - **Interval waktu, waiting time**
- applies to occurrences of some event over a **specified interval**
- random variable x is the number of occurrences of the event in an **interval**.
    - interval can be time, distance, area, volume, or some similar unit.

$$P(x) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad \text{where } e \approx 2.71828$$

Contoh Kasus

$$P(x) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad \text{where } e \approx 2.71828$$

- Jumlah pemakaian sebuah telepon **per jam** mengikuti distribusi Poisson dengan rata-rata = **9**. Berapakah probabilita akan dijumpai pemakaian telepon per jam sebanyak:
  - **2** kali?
  - Max. 3 kali?

$$P(2) = \frac{e^{-9} \cdot 9^2}{2!} = \frac{e^{-9} \cdot 81}{2} = 0{,}0049$$

$$P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

$$= \frac{e^{-9} \cdot 9^0}{0!} + \frac{e^{-9} \cdot 9^1}{1!} + \frac{e^{-9} \cdot 9^2}{2!} + \frac{e^{-9} \cdot 9^3}{3!}$$

$$= 0{,}0212$$

Try it with R

Binomial vs Poisson

- The Poisson distribution differs from the binomial distribution in these fundamental ways:
    - The binomial distribution is affected by the sample size n and the probability p, whereas the Poisson distribution is affected only by the mean μ.
    - In a binomial distribution the possible values of the random variable x are 0, 1, . . . n, but a Poisson distribution has possible x values of 0, 1, 2, . . . , with no upper limit.
- For binomial: when **n is large** and **p is small**, we can use **Poisson** as a proxy of **binomial**.
    - Konvensi: **n besar → 30**, **p kecil → < 0.1**
    - **μ = n*p**

Contoh

Probabilita kerusakan sebuah paku khusus pada permukaan sayap pesawat terbang baru adalah **0,001**. Ada **4,000** paku pada sayap pesawat terbang tersebut. Berapakah probabilita bahwa akan terdapat **tidak lebih dari 2 paku yang rusak**?

Answer:

Ini sebenarnya merupakan kasus **binomial** (**rusak/tidak rusak**). Tapi karena n besar dan p kecil, kita bisa gunakan pendekatan poisson, dimana **µ** = **np** = **4**.

Try it with R

$$P(x \leq 2) = P(x=0) + P(x=1) + P(x=2)$$

$$= \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!} + \frac{e^{-4} \cdot 4^2}{2!}$$

$$= 0,2381033$$

Refreshing Questions

1. Jika probabilita memperoleh bayi laki-laki dari suatu kelahiran adalah 0.5. Berapa besar probabilita 2 anak laki-laki dari 3 kelahiran?
   Answer:

   $$\boxed{{}^{n}C_{x} * p^{x} * q^{n-x}}$$

   $$P(x=2) = {}^{3}C_{2} * (0,5)^{2} (0,5)$$
   $$= 3 \times 0,125 = 0,375$$

Refreshing Questions

2. Rata-rata banyaknya partikel radioaktif yg melewati suatu penghitung **selama 1 milidetik** dalam suatu percobaan di laboratorium adalah **4**. Berapakah peluang 6 partikel melewati penghitung itu dalam 1 milidetik tertentu?

   Answer:

$$P(x=6) = \frac{e^{-4} \cdot 4^{6}}{6!} = 0{,}1042$$

$$P(x) = \frac{e^{-\mu} \cdot \mu^{x}}{x!} \quad \text{where } e \approx 2.71828$$

Refreshing Questions

3. Dari random variables berikut, mana yang diskret dan mana yang kontinu?
    a. Banyaknya kecelakaan mobil per tahun di Jakarta → diskret
    b. Lamanya waktu pertandingan sepak bola → kontinu
    c. Banyaknya produksi susu seekor sapi betina selama setahun. → kontinu
    d. Banyaknya telur yang dihasilkan setiap bulan oleh seekor ayam bentina → diskret
    e. Banyaknya SIM yang dikeluarkan tiap bulan di suatu kota tertentu → diskret
    f. Berat padi yang dihasilkan per hektar. → kontinu

Refreshing Questions

4. Tukang cuci mobil dibayar berdasarkan banyaknya mobil yang dia cuci. Misalkan bahwa penerimaannya dalam sehari, dalam ribuan rupiah, 7, 9, 11, 13, 15, atau 17 dengan peluang masing2: **1/12, 1/12, ¼, ¼, ⅙, dan ⅙, cari nilai harapan penghasilannya dalam sehari. Apa interpretasinya?**

x	P(x)	x.P(x)
7	0.08333333333	0.5833333333
9	0.08333333333	0.75
11	0.25	2.75
13	0.25	3.25
15	0.1666666667	2.5
17	0.1666666667	2.833333333
sum	1	12.66666667

Refreshing Questions

5. Peluang seseorang meninggal karena infeksi pernafasan adalah 0.002. Carilah peluang dari 2000 orang yang terinfeksi, kurang dari 5 orang diantaranya, meninggal.

p=0.002, q=0.998

n=2000 x=5, ➜ didekati dg poisson ➜ **μ = n*p = 2000*0.002 = 4**

Date

Binomial

Sampel besar, peluang kecil → dekati dengan Poisson

$$\mu = n.p = 2000 * 0{,}002 = 4$$

$$P(x<5) = \sum_{x=0}^{4} \frac{e^{-\mu} \cdot \mu^{x}}{x!}$$

$$= \frac{e^{-4} \cdot 4^{0}}{0!} + \frac{e^{-4} \cdot 4^{1}}{1!} + \frac{e^{-4} \cdot 4^{2}}{2!} + \frac{e^{-4} \cdot 4^{3}}{3!} + \frac{e^{-4} \cdot 4^{4}}{4!}$$

$$= 0{,}6288369$$

Refreshing Questions

6. In a certain town, 40% of the eligible voters prefer candidate A, 10% prefer candidate B, and the remaining 50% have no preference. You randomly sample 10 eligible voters. What is the probability that 4 will prefer candidate A, 1 will prefer candidate B, and the remaining 5 will have no preference?

Multinomial

$$\frac{10!}{4!\,1!\,5!}\,(0{,}4)^4 \cdot (0{,}1)^1 \cdot (0{,}5)^5$$

$$= 0{,}1008$$

$$\frac{n!}{x_1!\,x_2!\,x_3!\cdots x_k!}\,p_1^{x_1}\,p_2^{x_2}\cdots p_k^{x_k}$$

Thank You