

# Beberapa Distribusi Kontinu dan Diskrit

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

*aliakbars@live.com*

March 30, 2018

# Selayang Pandang

- 1 Ulasan
- 2 Distribusi Uniform
- 3 Distribusi Gaussian
  - Univariate Gaussian
  - Maximum Likelihood Estimation
  - Multivariate Gaussian
- 4 Distribusi Beta, Bernoulli, dan Binomial

# Ulasan

## Minggu lalu...

- Peubah acak, ruang sampel, kejadian  
Jika dilempar sebuah dadu,  $p(X = 5) = \dots$

## Minggu lalu...

- Peubah acak, ruang sampel, kejadian  
Jika dilempar sebuah dadu,  $p(X = 5) = \dots$
- Ekspektasi dan variansi  
 $\mathbb{E}[X] = \dots$  dan  $\text{Var}[X] = \dots$

## Minggu lalu...

- Peubah acak, ruang sampel, kejadian  
Jika dilempar sebuah dadu,  $p(X = 5) = \dots$
- Ekspektasi dan variansi  
 $\mathbb{E}[X] = \dots$  dan  $\text{Var}[X] = \dots$
- Marginal probability, conditional probability  
sum rule, product rule, chain rule

## Minggu lalu...

- Peubah acak, ruang sampel, kejadian  
Jika dilempar sebuah dadu,  $p(X = 5) = \dots$
- Ekspektasi dan variansi  
 $\mathbb{E}[X] = \dots$  dan  $\text{Var}[X] = \dots$
- Marginal probability, conditional probability  
**sum rule, product rule, chain rule**
- Bayes' rule  
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

# Ekspektasi Kontinu dan Variansi

## Ekspektasi

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot p(x) dx$$



# Ekspektasi Kontinu dan Variansi

## Ekspektasi

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot p(x) dx$$

## Variansi

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Tunjukkan!

Sudah coba soal latihan?

Sudah enroll ke e-learning?

# Distribusi Uniform

## Contoh



The image shows the Spotify interface for Taylor Swift's album 'reputation'. The album cover features a black and white photo of Taylor Swift with a newspaper clipping overlay. The title 'reputation' is in a large, bold, white font. Below the title, it says 'Par Taylor Swift' and '2017 • 15 titres, 55 min'. There are buttons for 'LIRE' (green) and 'SAUVEGARDER' (white on dark background). Below the album information is a tracklist table.

| # | TITLE                         | Duration |
|---|-------------------------------|----------|
| 1 | ...Ready For It?              | 3:28     |
| 2 | End Game - Ed Sheeran, Future | 4:05     |
| 3 | I Did Something Bad           | 3:58     |
| 4 | Don't Blame Me                | 3:56     |
| 5 | Delicate                      | 3:52     |
| 6 | Look What You Made Me Do      | 3:32     |
| 7 | So It Goes...                 | 3:48     |
| 8 | Gorgeous                      | 3:30     |
| 9 | Getaway Car                   | 3:54     |

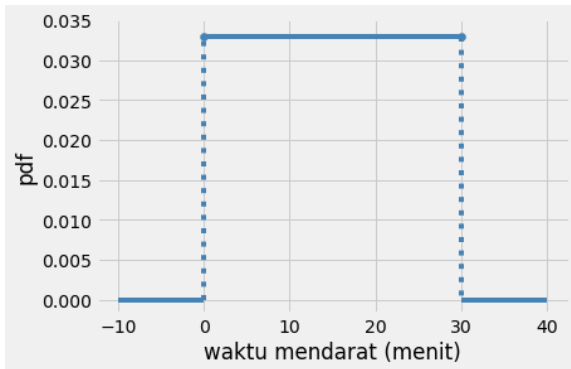
**Gambar:** Durasi lagu-lagu Taylor Swift dalam album Reputation cenderung terdistribusi seragam

# Peubah Acak Seragam

- Semua bernilai sama peluangnya dalam interval tertentu
- Dituliskan sebagai  $X \sim Uni(\alpha, \beta)$
- Bisa berupa diskrit maupun kontinu
- PDF:

$$p(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{if } x \in [\alpha, \beta] \\ 0, & \text{otherwise} \end{cases}$$

## Contoh



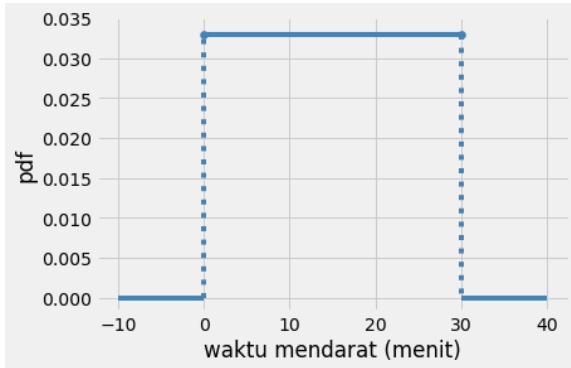
**Gambar:** Berapa probabilitas pesawat mendarat ada di antara 25-30 menit?

Mengapa PDF di satu titik bisa  $> 1$ ?



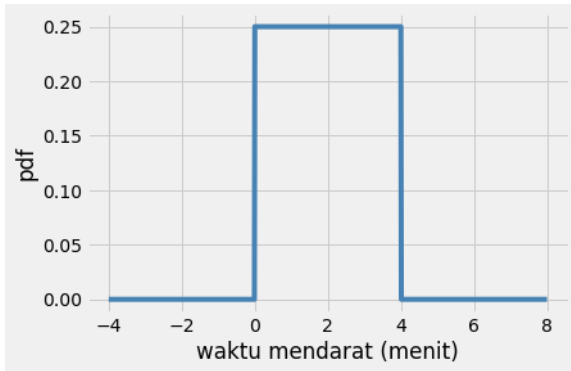
## Perhatikan!

Ingat bahwa  $\int_{-\infty}^{\infty} p(x)dx = 1$  sehingga luas area di bawah (persegi panjang) harus bernilai 1.

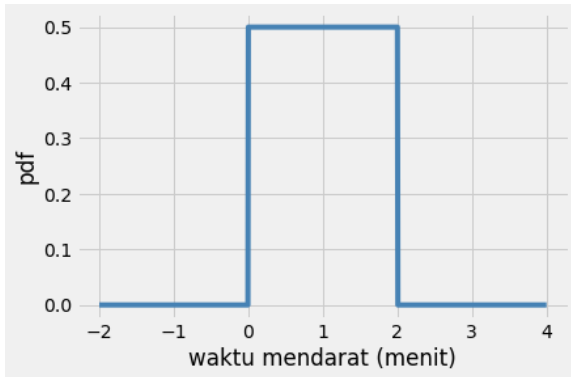


Apa yang terjadi kalau intervalnya kita perkecil?

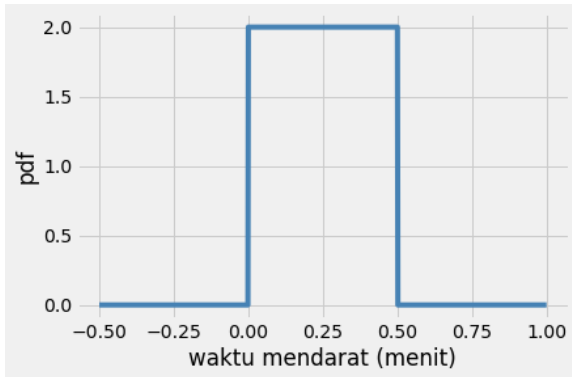
Perhatikan!



Perhatikan!



Perhatikan!



# Ekspektasi dan Variansi

## Ekspektasi

$$\mathbb{E}[X] = \frac{\alpha + \beta}{2}$$

## Variansi

$$\text{Var}[X] = \frac{(\beta - \alpha)^2}{12}$$

# Distribusi Gaussian

# Distribusi Gaussian/Normal

- Salah satu yang paling sering muncul untuk variabel kontinu
- Berhubungan dengan central limit theorem
- Dituliskan sebagai  $X \sim \mathcal{N}(\mu, \sigma^2)$



## Apa itu “terdistribusi normal”?

- Dapat ditemukan dalam berbagai fenomena di alam, e.g. tinggi badan, berat badan, ...

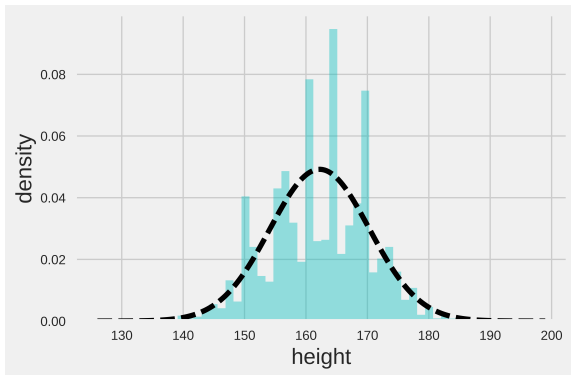
## Apa itu “terdistribusi normal”?

- Dapat ditemukan dalam berbagai fenomena di alam, e.g. tinggi badan, berat badan, ...
- Jumlah dari berbagai peubah acak yang independen

## Apa itu “terdistribusi normal”?

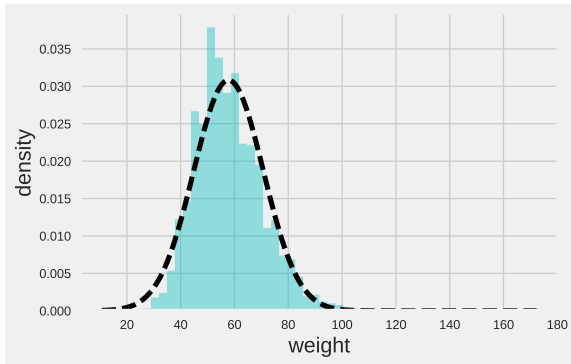
- Dapat ditemukan dalam berbagai fenomena di alam, e.g. tinggi badan, berat badan, ...
- Jumlah dari berbagai peubah acak yang independen
- Dengan jumlah sampel yang cukup, bisa menggambarkan populasi dengan baik

## Contoh



Gambar: Hasil “pengukuran” tinggi badan

## Contoh



Gambar: Hasil pengukuran berat badan

## Fakta

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

## PDF

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

## Ekspektasi

$$\mathbb{E}[X] = \mu$$

## Variansi

$$\text{Var}[X] = \sigma^2$$

## CDF

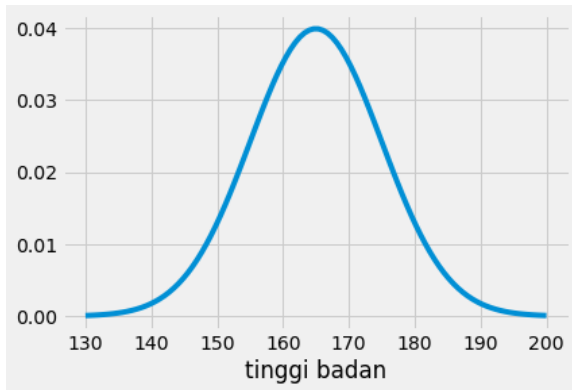
- Karena kita berurusan dengan distribusi kontinu, kita perlu **cumulative density function**
- CDF:

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \int_{-\infty}^x f(x)dx$$

- e.g. Berapa peluangnya untuk mendapatkan orang dengan tinggi badan antara 150 dan 160?

## Solusi

$$p(150 < X < 160) = F(160) - F(150)$$

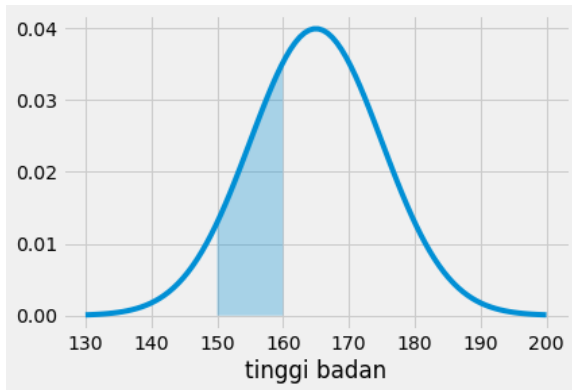


Gambar: Distribusi tinggi badan



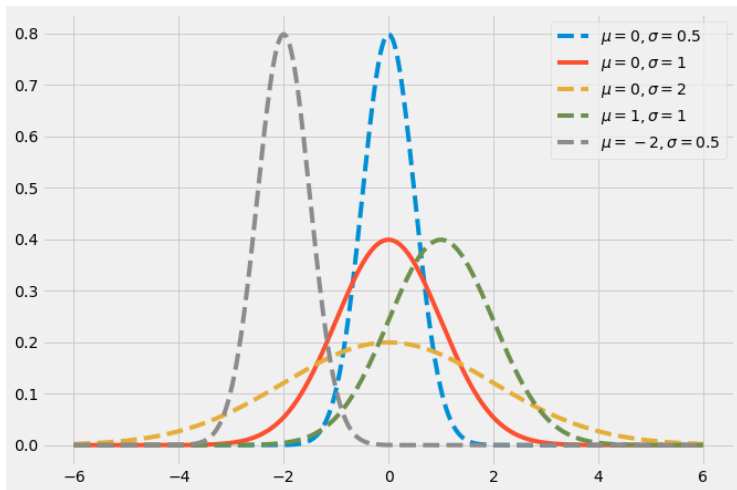
## Solusi

$$p(150 < X < 160) = F(160) - F(150)$$



Gambar: Distribusi tinggi badan

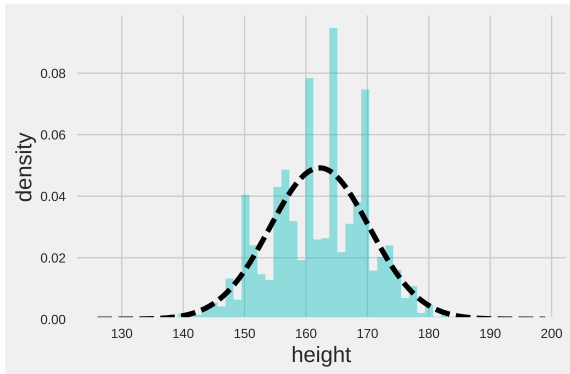
Perhatikan bahwa...



Gambar: Distribusi Gaussian dengan berbagai nilai parameter

Jika semua nilai hasil observasi sama, bagaimana grafiknya?

## Contoh



Gambar: Bagaimana cara mendapatkan distribusi Gaussian<sup>1</sup>?

<sup>1</sup>digambarkan dengan garis putus-putus

## Likelihood

- $p(\mathcal{D}|\mathcal{M})$ , yaitu probabilitas melihat data  $\mathcal{D}$  jika diberikan distribusi (atau model)  $\mathcal{M}$
- Merupakan hasil perkalian dari peluang yang menghasilkan tiap titik dalam data

$$L(\mathcal{M}) = p(\mathcal{D}|\mathcal{M}) = \prod_{i=1}^N p(x_i|\mathcal{M})$$

- Kita menggunakan asumsi *independent and identically distributed* (i.i.d.)
- Datanya tetap, modelnya (parameternya) yang dapat berubah

# Maximum Likelihood Estimation

- Coba berbagai model  $\mathcal{M}$  yang dapat memaksimalkan nilai *likelihood*, i.e. *maximum likelihood estimation*
- Dalam kasus distribusi Gaussian

$$L(\mathcal{M}) = p(X|\mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

- Atur  $\gamma = 1/\sigma^2$ , lalu cari titik optimumnya.
- Bagaimana?

## Beberapa trik...

- Terapkan fungsi logaritma
- Cari turunan parsial pertama
- Atur sama dengan nol

## Maximum Likelihood Estimation

$$\log p(X|\mu, \gamma) = -\frac{1}{2} \sum_{i=1}^N \gamma (x_i - \mu)^2 - \frac{N}{2} \log(2\pi) + \frac{N}{2} \log \gamma$$

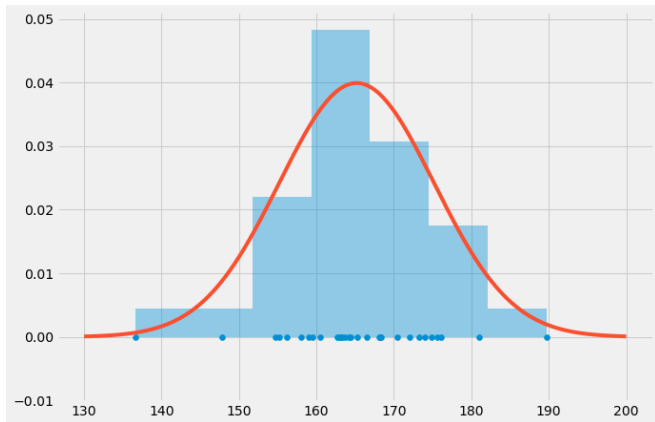
$$\frac{\partial p(X|\mu, \gamma)}{\partial \mu} = \gamma \sum_{i=1}^N (x_i - \mu)$$

$$\frac{\partial p(X|\mu, \gamma)}{\partial \gamma} = -\frac{1}{2} \sum_{i=1}^N (x_i - \mu)^2 + \frac{N}{2\gamma}$$

sehingga pada titik maksimum:  $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$  dan  $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$ . Terlihat familiar?



## Grafik MLE



Gambar: Gaussian MLE dari 30 objek dalam data

# Multivariate Gaussian

- Vektor  $\mathbf{x}$  adalah multivariate Gaussian jika untuk *mean*  $\mu$  dan *covariance matrix*  $\Sigma$ , nilainya terdistribusi menurut

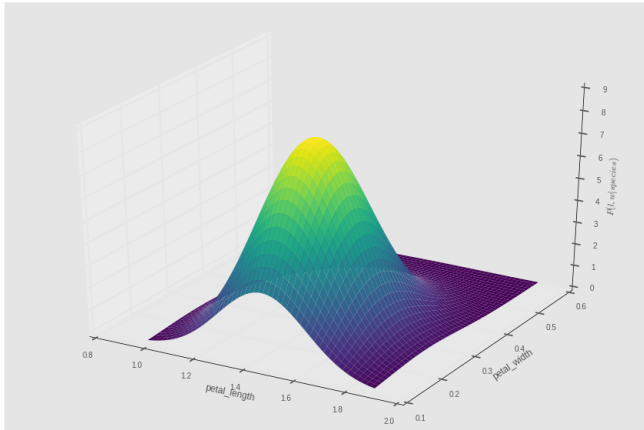
$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{|(2\pi)\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

- Univariate Gaussian adalah kasus khusus dari distribusi ini
- $\Sigma$  adalah *covariance matrix*, i.e. setiap elemen  $\sigma_{ij} = \text{Cov}(X_i, X_j)$  dengan

$$\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

- $\Sigma$  harus simetris

# Bivariate Gaussian

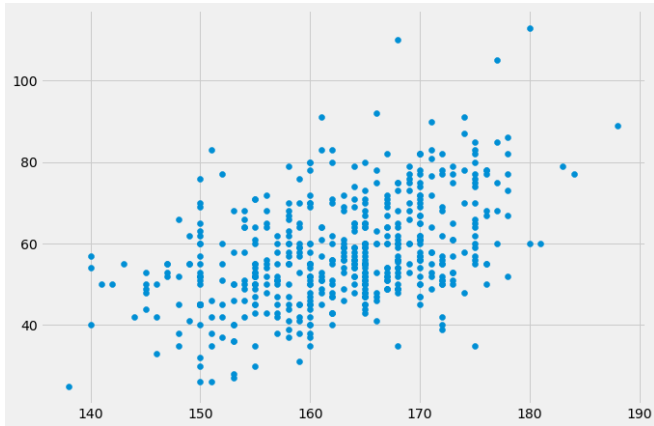


**Gambar:** Multivariate Gaussian dengan dua variabel yang dibuat dalam tiga dimensi

# Maximum Likelihood

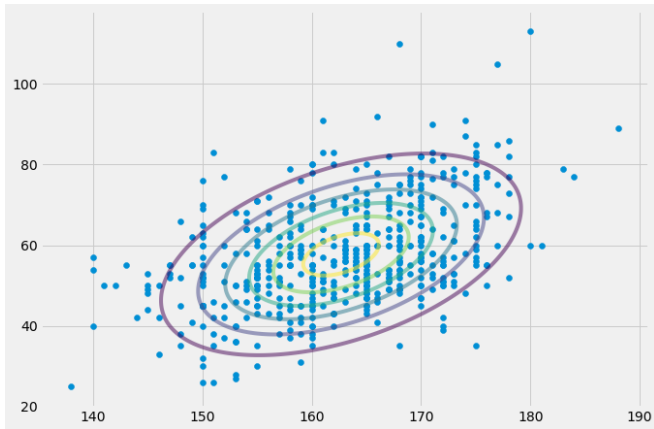
- Sama dengan kasus univariate
- $\mu = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$
- $\Sigma = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$

## Grafik MLE



Gambar: Data tinggi vs berat badan

## Grafik MLE



Gambar: MLE dari data

## Poin Penting

- Jumlah dari Gaussian RVs adalah Gaussian
- Model yang terdiri dari kombinasi linear Gaussian akan memiliki *joint distribution* berupa Gaussian
- Jika  $p(x, y)$  adalah multivariate Gaussian, maka  $p(x)$  maupun  $p(y)$  serta  $p(x|y)$  dan  $p(y|x)$  adalah Gaussian

# Distribusi Beta, Bernoulli, dan Binomial



# Distribusi Bernoulli

- $X$  adalah RV yang bisa bernilai 0 atau 1
- Jika  $p(X = 1|\theta) = \theta$  dan mengakibatkan  $p(X = 0|\theta) = 1 - \theta$ , maka
- $X$  mengikuti distribusi Bernoulli

# Model Bernoulli

## Example

Data: 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1.

Tiga hipotesis:

- $\mathcal{M} = 1$  - dari koin setimbang  $1 = H, 0 = T$
- $\mathcal{M} = 2$  - dari lemparan dadu  $1 = 1, 0 = 2, 3, 4, 5, 6$
- $\mathcal{M} = 3$  - dari koin yang keduanya muka  $1 = H, 0 = T$

## Model Bernoulli

### Example

Data: 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1.

*Likelihood of data.* Jika  $N_1$  = jumlah 1,  $N_0$  = jumlah 0, dengan  $N = N_0 + N_1$ :

$$\prod_{i=1}^N p(x_i|\mathcal{M}) = p(1|\mathcal{M})^{N_1} p(0|\mathcal{M})^{N_0}$$

maka

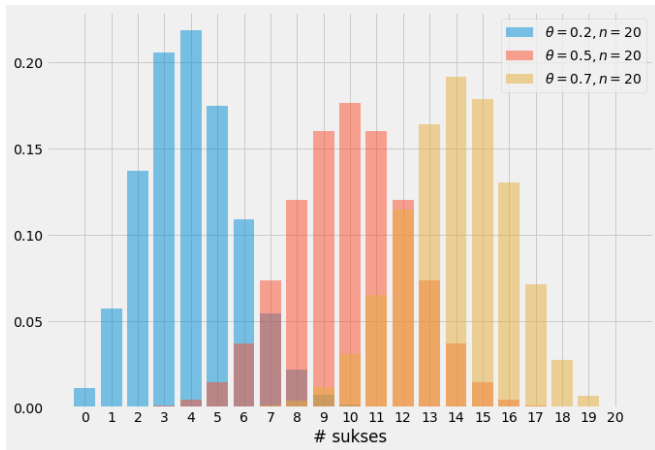
- $\mathcal{M} = 1 : L(\mathcal{M}) = 0.5^{20} = 9.5 \times 10^{-7}$
- $\mathcal{M} = 2 : L(\mathcal{M}) = \left(\frac{1}{6}\right)^9 \left(\frac{5}{6}\right)^{11} = 1.3 \times 10^{-8}$
- $\mathcal{M} = 3 : L(\mathcal{M}) = 1^9 0^{11} = 0$

Coba cari MLE-nya!

# Distribusi Binomial

- Didapatkan dari  $n$  percobaan Bernoulli independen
- Distribusinya menunjukkan jumlah kemunculan 1
- $X$  dapat bernilai  $0, 1, 2, \dots, n$
- Jika  $p(X = r|\theta) = \binom{n}{r}\theta^r(1 - \theta)^{(n-r)}$ ,
- maka  $X$  mengikuti distribusi Binomial

## Grafik PMF



Gambar: Probability mass function (PMF) dari distribusi Binomial

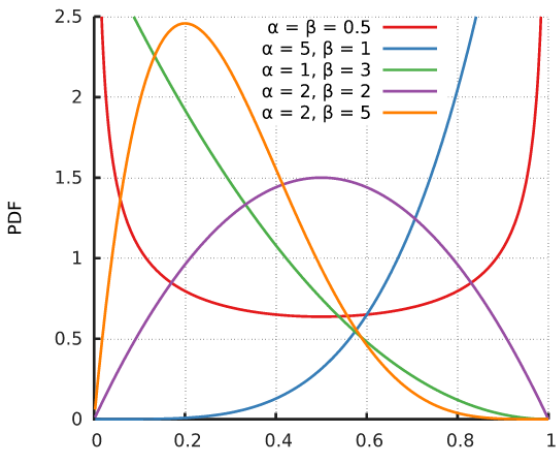
## Distribusi Beta

- Merepresentasikan keluaran dari persentase atau proporsi
- Berguna untuk dipakai sebagai *prior probability*
- Jika  $X$  adalah RV yang bernilai  $x \in [0, 1]$ ,
- dan

$$p(X = x|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{(a-1)}(1-x)^{(b-1)}$$

- maka  $X \sim \text{Beta}(a, b)$ .

## Grafik PDF



Gambar: PDF dari distribusi Binomial



- Distribusi uniform
- Distribusi normal/Gaussian
- Maximum Likelihood Estimation
- Distribusi Bernoulli, Binomial, dan Beta

## Pertemuan Berikutnya

- Bayes classifier
- Naïve Bayes
- Conditional independence

Office hours minggu ini di  
hari Rabu, 08.00-09.00

# Referensi



Chris Piech (Jul. 2017)

The Normal Distribution

<http://web.stanford.edu/class/cs109/>



Chris Williams (Sep. 2015)

Probability - Machine Learning and Pattern Recognition

<https://www.inf.ed.ac.uk/teaching/courses/mlpr/2015/>



Chris Williams (Sep. 2015)

The Gaussian Distribution - Machine Learning and Pattern Recognition

<https://www.inf.ed.ac.uk/teaching/courses/mlpr/2015/>

Terima kasih