Beberapa Distribusi Kontinu

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

March 5, 2020

Selayang Pandang

Ulasan

Ekspektasi Kontinu dan Variansi

Ekspektasi

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

Variansi

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

Tunjukkan!

Sudah enroll ke e-learning?

Distribusi Uniform

Replenishment Lead Time

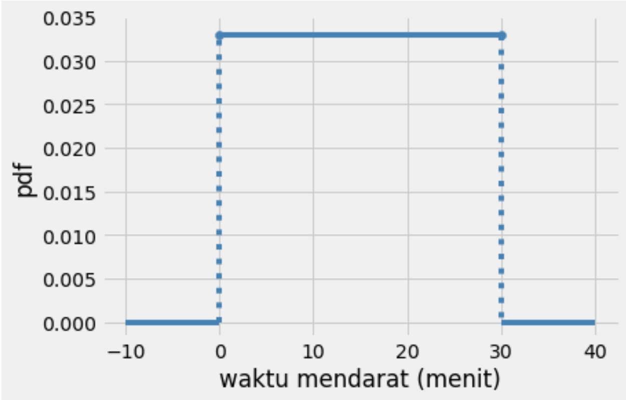
Gambar: Waktu yang dibutuhkan dari pemesanan hingga masuk ke gudang bisa diasumsikan terdistribusi uniform (Chopra et al., 2004)

Peubah Acak Seragam

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

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

Contoh

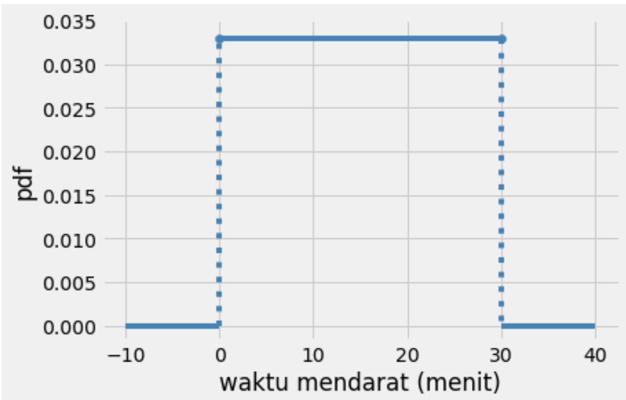


Gambar: Berapa probabilitas pesawat mendarat ada di antara 25-30 menit?

Mengapa PDF di satu titik bisa > 1?

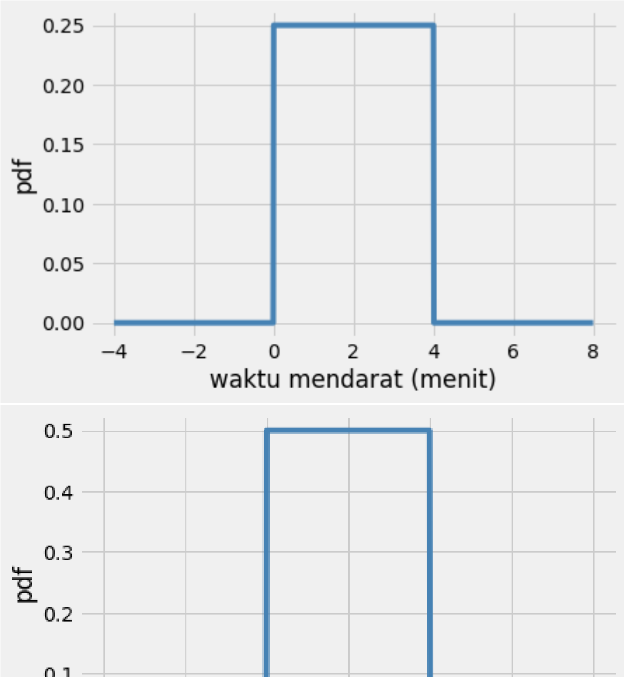
Perhatikan!

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



Apa yang terjadi kalau intervalnya kita perkecil?

Perhatikan!



Ekspektasi dan Variansi

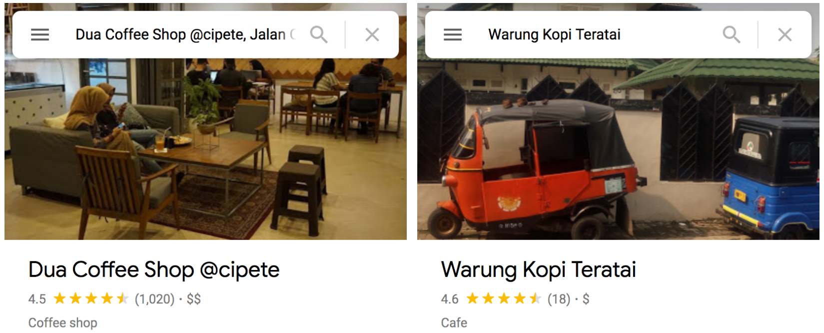
Ekspektasi

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

Variansi

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

Distribusi Beta



Gambar: Mana yang lebih Anda percaya?

Kita ingin memasukkan unsur

ketidakpastian

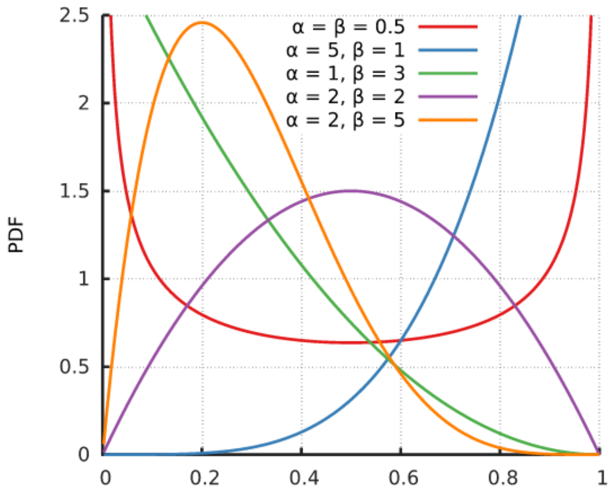
Distribusi Beta

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

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

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

Grafik PDF



Gambar: PDF dari distribusi Binomial

Ekspektasi dan Variansi

Ekspektasi

$$\mathbb{E}[X] = \frac{a}{a+b}$$

Variansi

$$Var[X] = \frac{ab}{(a+b)^2(a+b+1)}$$

Example

Kafe Romeo diberikan nilai 5, 4, 2, 1, 4, 5, 5, 5, 4, 4, 5, 5, 5, 5, 5. Kafe Juliet diberikan nilai 5, 5, 4. Mana yang lebih baik?

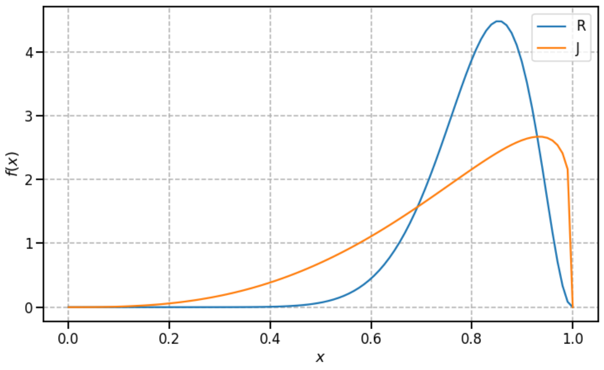
Solution

Jika $X \sim Beta(a, b)$, maka $X \in [0, 1]$. Jadi, ubah skala nilainya terlebih dahulu, e.g. $2 \to 0.4$.

$$a = 1 + S$$
$$b = 1 + N - S$$

dengan N adalah jumlah orang yang memberi nilai, dan S adalah jumlah nilai yang telah diubah skalanya.

Solusi



$$X_R \sim Beta(a_R, b_R) = Beta(13.8, 3.2)$$
$$X_J \sim Beta(a_J, b_J) = Beta(3.8, 1.2)$$

Solusi (lanjutan)

Dengan asumsi central limit theorem, kita dapat menghampiri distribusi Beta dengan Gaussian. Lalu, anggaplah kita mau melihat kemungkinan terburuk. Dengan demikian, nilai di kuantil .5 dapat dihitung dengan (95% *least plausible value*):

$$score = \mu - 1.65\sigma$$

atau

$$score = \frac{a}{a+b} - 1.65\sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$$

sehingga

$$score_R = 3.30$$
$$score_J = 2.36$$

Distribusi Gaussian

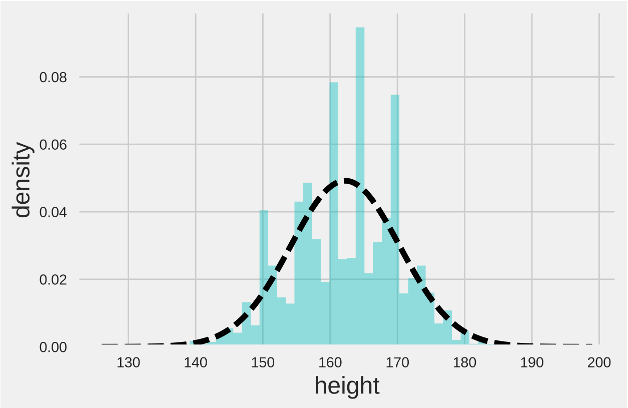
Distribusi Gaussian/Normal

- 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"?

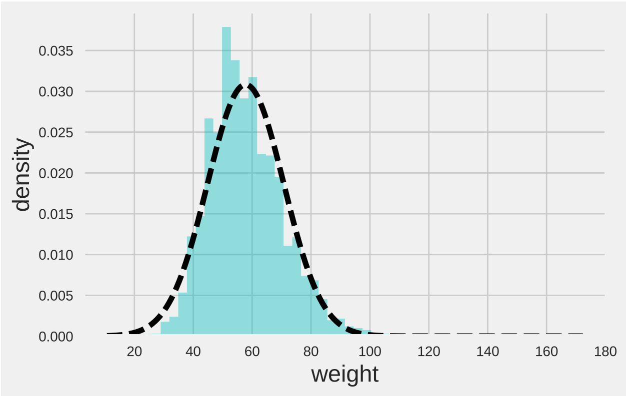
- Dapat ditemukan dalam berbagai fenomena di alam, e.g. tinggi badan, berat badan, ...
- Jumlah dari berbagai peubah acak yang independen
- Dengan jumlah sampel yang cukup, bisa menggambarkan populasi dengan baik

Contoh



Gambar: Hasil "pengukuran" tinggi badan

Contoh



Gambar: Hasil pengukuran berat badan

Fakta

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

PDF

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

Ekspektasi

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

Variansi

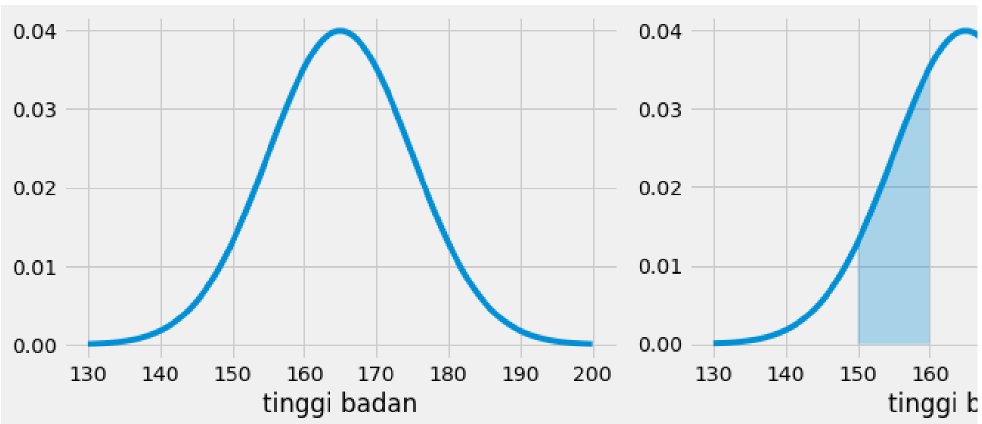
$Var[X] = \sigma^2$

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?

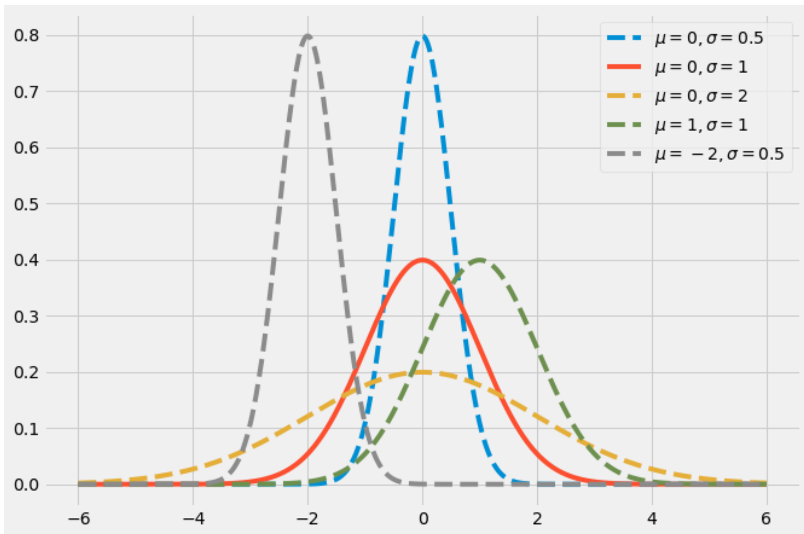
Solusi

$$p(150 \leq X \leq 160) = F(160) - F(150)$$



Gambar: Distribusi tinggi badan

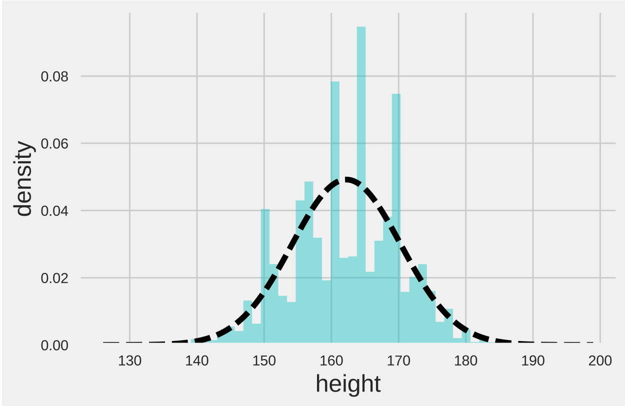
Perhatikan bahwa...



Gambar: Distribusi Gaussian dengan berbagai nilai parameter

Contoh

Jika semua nilai hasil observasi sama, bagaimana grafiknya?



Gambar: Bagaimana cara mendapatkan distribusi Gaussian[1]?

[1] digambarkan dengan garis putus-putus

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.

Maximum Likelihood Estimation

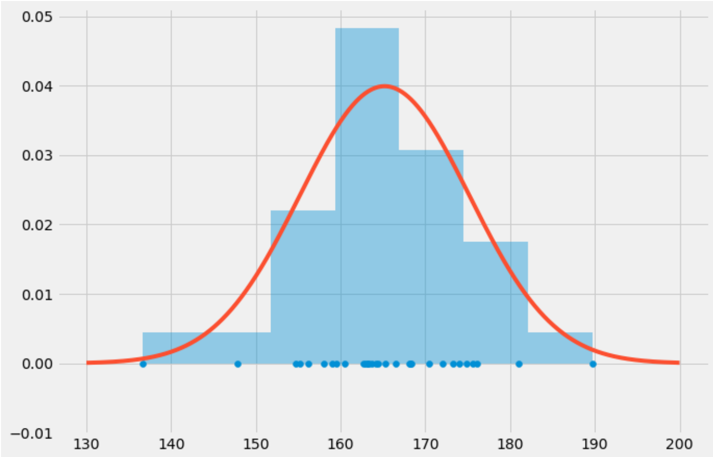
$$\log p(X|\mu,\gamma) = -\frac{1}{2}\sum_{i=1}^{N}\gamma(x_i-\mu)^2 - \frac{N}{2}\log(2\pi) + \frac{N}{2}\log\gamma$$

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

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

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* μ dan *covariance matrix* Σ, nilainya terdistribusi menurut

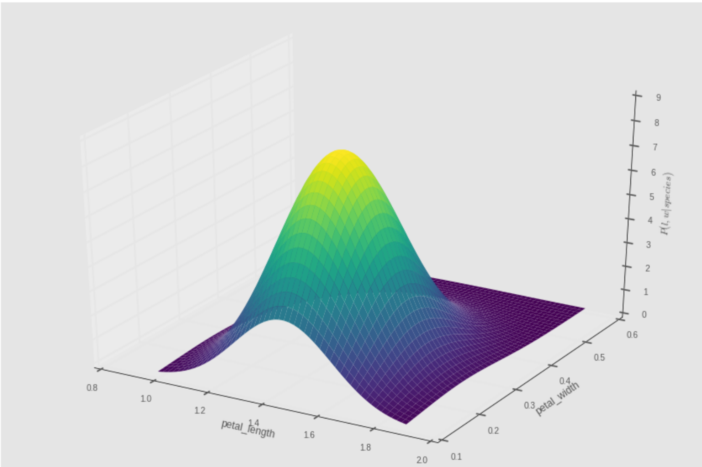
$$f(\mathbf{x}|\mu, \Sigma) = \frac{1}{|(2\pi)\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

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

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

- Σ harus simetris

Bivariate Gaussian

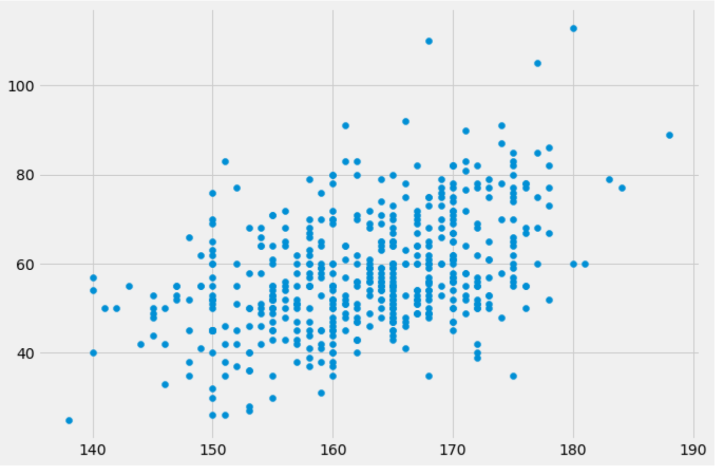


Gambar: Multivariate Gaussian dengan dua variabel yang dibuat dalam tiga dimensi

Maximum Likelihood

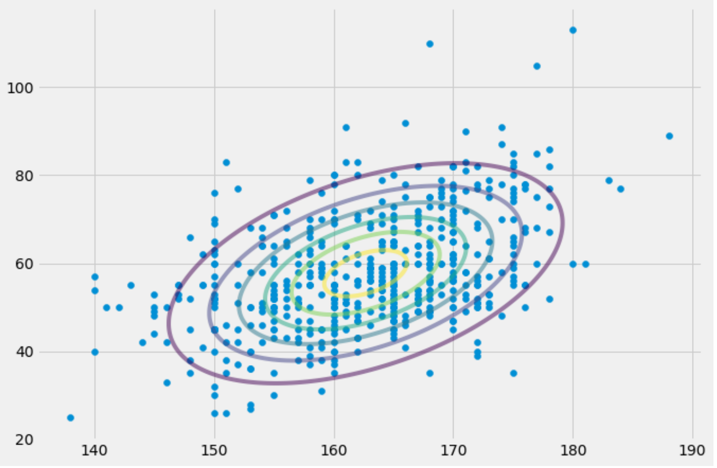
- Sama dengan kasus univariate
- $\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$
- $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$

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

Ikhtisar

- Distribusi uniform
- Distribusi Beta
- Distribusi normal/Gaussian

Pertemuan Berikutnya

- Bayes classifier
- Naïve Bayes
- Conditional independence

Referensi

Will Monroe (Jul. 2017)
The Normal Distribution
http://web.stanford.edu/class/archive/cs/cs109/cs109.1178/lectureHandouts/110-normal-distribution.pdf

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