Essential Mathematics for Global Leaders I

Spring 2019

Statistics

Lecture 4: 2019 May 20 – May 27

Xavier DAHAN
Ochanomizu Graduate Leading Promotion Center

Office:理学部2号館503
mail: dahan.xavier@ocha.ac.jp

Where are we ? Today's plan

PART I. Notions of Probability 必要な確率論

2. Random variable 確率変数

2.1 Discrete random variable 離散確率変数
2.2 Some important distributions 幾つか大事な分布
2.3 Operations on random variables 確率変数への作用
2.4 Expected value 期待値
2.5 Variance 分散

Lecture 3

2.6 Continuous random variable 連続確率変数

2.7 Some important continuous random variables 幾つか大事な連続確率変数の分布

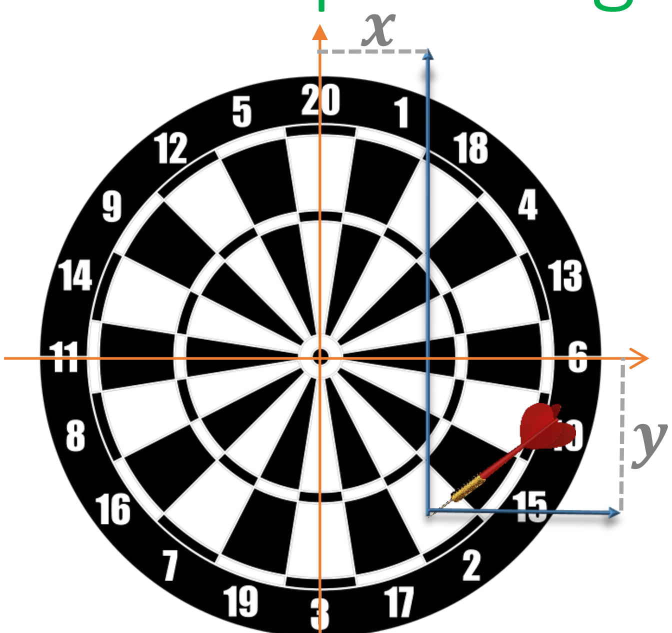
2.8 Expected value and variance (continuous case) 期待値と分散（連続の場合）

Chapter 2: Random variable

2.6 continuous random variable

2.6 連続確率変数

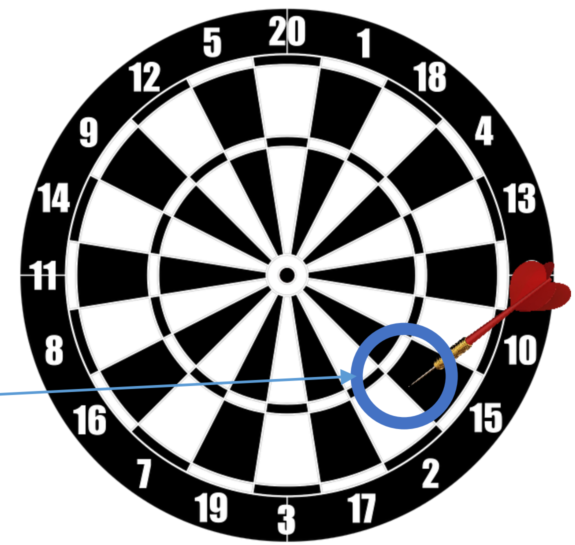
Example: target and darts. 的とダーツ



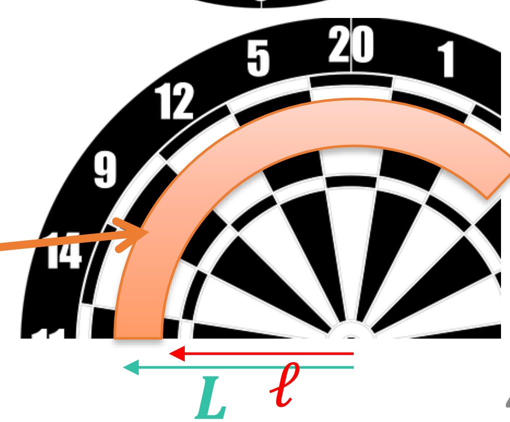
- Experiment: throw a darts to the target.
  事件：的にダーツを投げる
- Sample space: position (x, y) on the target.
  標本空間：的での立ち位置の座標(x, y)
- In this example, the sample space Ω is not discrete (not countable). It is continuous. 標本空間は離散ではない（可付番ではない）連続という。

Probability of a continuous sample space

- What is the probability that a dart lands at a given position (a, b) on the target?
  ダーツは正確に立ち位置(a,b)に着く確率は何か？
- $P(\textit{ lands at } (a,b) \textit{ exactly}) = 0.$

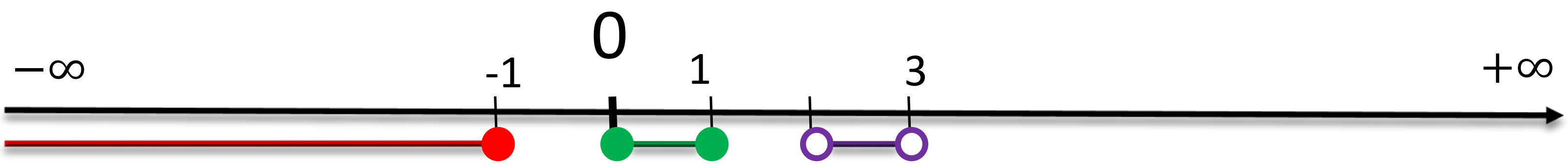


- Needs to consider landing regions
  着く領域の方を考える。
- $P(\textit{lands in blue region}) = \dfrac{\textit{area of blue region}}{\textit{area of the target}}$
- **Random variable**: measure the distance to the center.原点から距離
  $X: \Omega \to \mathbb{R}, \qquad (x, y) \mapsto \sqrt{x^2 + y^2}$
- $P(\ell \leq X \leq L) = \dfrac{\textit{area in orange}}{\textit{area of target}}$



Probability and integration: introduction
確率と積分：　あらすじ

- Needs to measure length and area etc. ☛ Integration
  面積と長さなどを計る必要。☛　積分
- $\mathbb{R}$ the continuous real line (連続実数直線)



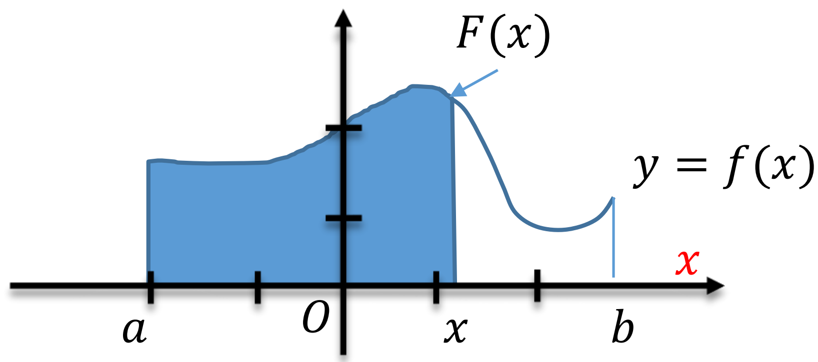
- Open interval 開区間: $(2,3)$
- Closed interval 閉区間 : $[0,1]$
- Semi-open interval 半区間：
  Open on left, closed on right: $(-\infty, -1]$

積分 (おさらい)　Integral (review)

Goal: Let $f: [a, b] \to \mathbb{R}$, $f > 0$ on $[a, b]$.
Define a function $F: [a, b] \to \mathbb{R}$ that measures the area between the curve and the x-axis
(関数のグラフとｘ-軸の間の面積を測る関数$F: [a, b] \to \mathbb{R}$を定義したい）。

Notation (記号):

$$F(x) = \int_a^x f(t)dt$$



Fundamental Theorem of Calculus (微分積分学の基本定理）：

$$\boldsymbol{F'(x) = f(x)}$$

$$\int_a^b f(t)\,dt = F(b) - F(a)$$

Probability density function (pdf)確率密度関数
Cumulative distribution function (cdf)累積分布関数

Definition (cdf): $F: \mathbb{R} \to \mathbb{R}$ is a cumulative distribution function (累積分布関数) if:

1. F is continuous 連続性: $\lim_{x \to c} F(x) = F(c)$.
2. F is non-decreasing 増加関数: $c \leq d \Rightarrow F(c) \leq F(d)$.
3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
4. $0 \leq F(c) \leq 1$.

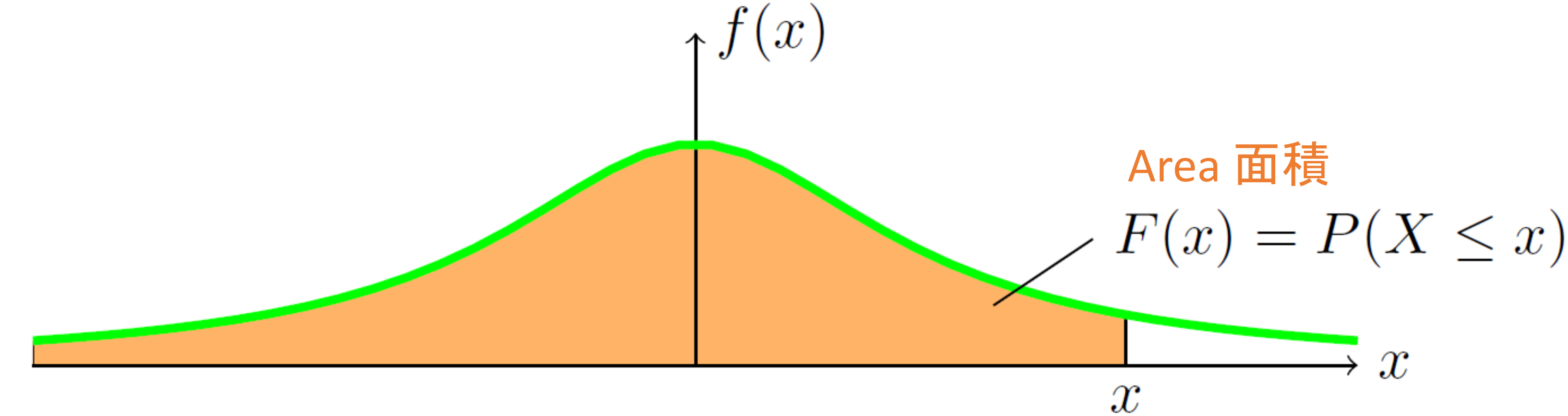
Definition: $X: \Omega \to \mathbb{R}$ is a **continuous random variable** if:

There exists a (probability density) function $f: \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that, for any $c \in \mathbb{R}$ the function $F(c)$ is equal to $P(X \leq c)$.

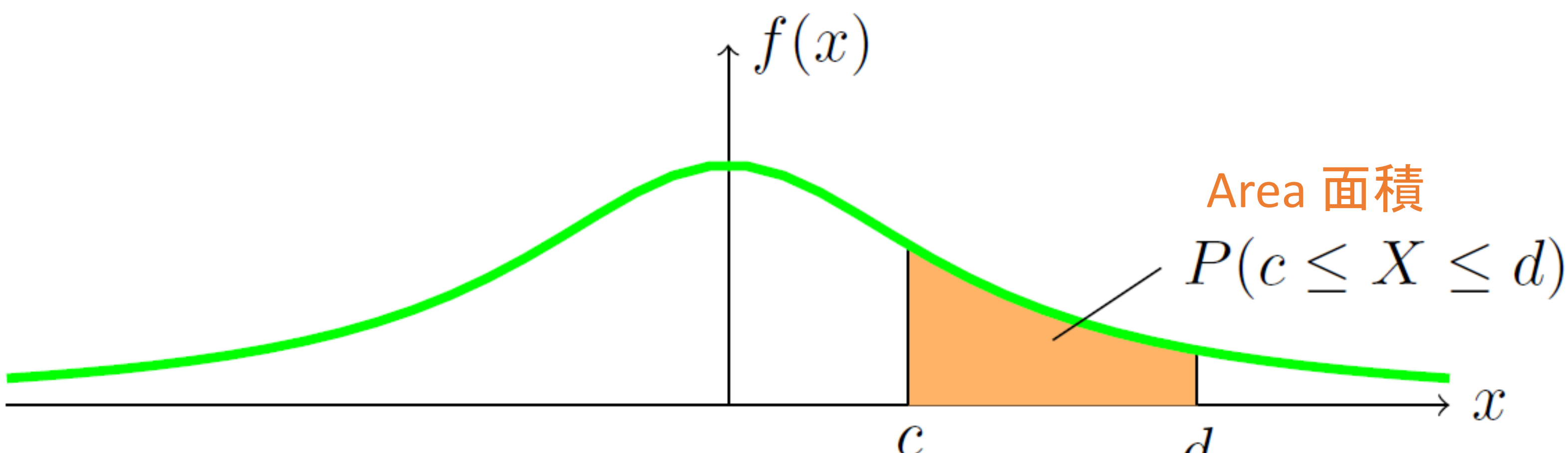
$$F(c) = P(X \leq c) := \int_{-\infty}^{c} f(x)\, dx$$

is a (continuous) cumulative distribution function (cdf).

Visualization　　可視化



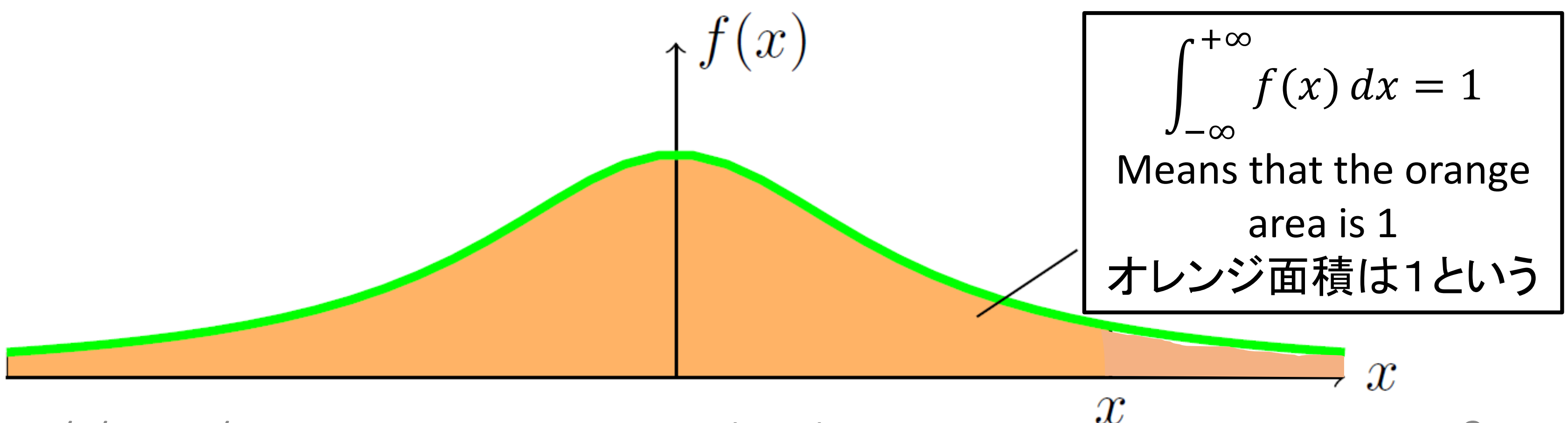
pdf $f(x)$ （確率密度関数） and **cdf** $\boldsymbol{F(x)}$ （累積分布関数）



pdf $f(x)$ （確率密度関数） and probability function （確率関数）

Properties of pdf （確率密度関数の性質）

- Positivity　正値性: $f(x) \geq 0$
- $F'(x) = f(x)$ (fundamental theorem of calculus)
  (微分積分の基本定理)
- $P(c \leq X \leq d) = \int_c^d f(x)\,dx = F(d) - F(c)$
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$　(pdf has mass 1　確率密度関数は質量１である).



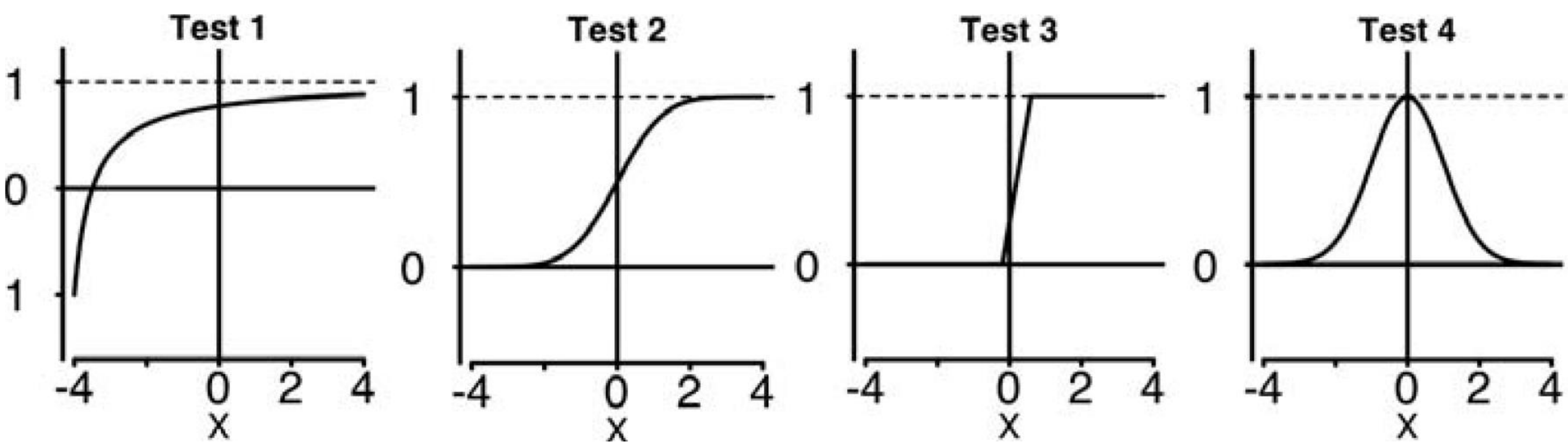
Comparison with discrete random variable
離散確率変数と比較

- Probability density function (pdf) $f(x)$ for a continuous random variable is the analogue of pmf (probability mass function) $p(x)$ for a discrete random variable, but
連続確率変数に対する確率密度関数$f(x)$は離散確率変数に対する確率質量関数$p(x)$に類似するといってもいい。ただ、

1. Pdf $f(x)$ is NOT a probability. You have to integrate to take the probability.
確率密度関数$f(x)$は確率ではない。確率を測定するために積分をとる必要がある。
2. In particular $f(x) > 1$ is not a problem.
特に、$f(x) > 1$があり得る。

Questions. Check understanding I

Which of the following are graphs of valid cumulative distribution functions?
下記グラフの中、なり得る累積分布がどれか。



- Answer: 2 and 3.

Questions. Check understanding II

1. Suppose X has range $[0,2]$ and pdf $f(x) = cx$.
（連続）確率変数Xの値域は$[0,2]$で、その密度関数$f(x) = cx$である（点数cは未知)
   a. What is the value of c.
   b. Compute the cdf $F(x)$.
   c. Compute $P(1 \leq X \leq 2)$.

Questions. Check understanding III

2. Suppose Y has range (值域) $[0, b]$ and cdf $F(y) = y^2/9$.
   a. What is b?
   b. Find the pdf of Y .

3. Suppose X is a continuous random variable.
   a. What is $P(a \le X \le a)$?
   b. What is $P(X = 0)$?
   c. Does $P(X = a) = 0$ mean X never equals a?

Chapter 2: Random variable

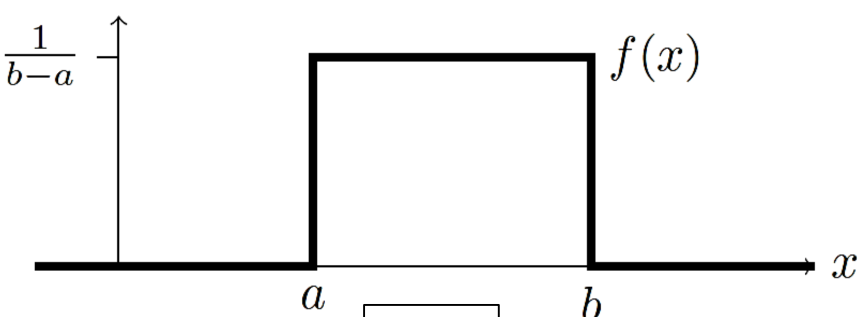
2.7 Some important continuous random variables （2.7 大事な連続確率変数）

- **Discrete （離散）**
  - $Bernoulli(p)$ ベルヌーイ
  - $Uniform(n)$ 一様分布
  - $Binomial(n,p)$ 二項分布
  - $Geometric(p)$ 幾何分布
- **Continuous （連続）**
  - $Uniform(a,b)$ 一様分布
  - $Normal(\mu,\sigma^2)$ 正規分布
  - $Exponential(\lambda)$ 指数分布

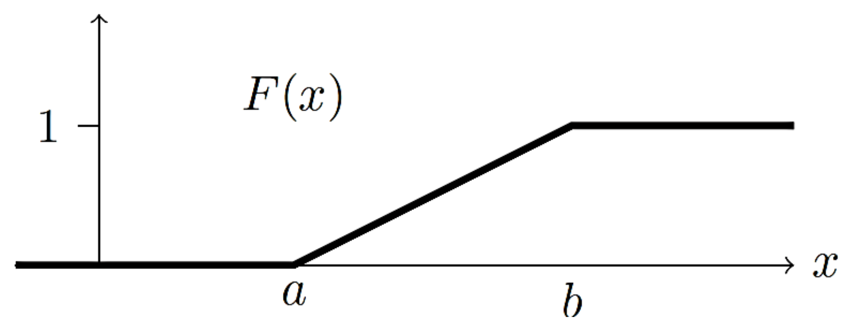
Uniform distribution (一様分布）

- $X \sim Uniform(a,b)$ $(or\ X \sim U(a,b))$
  1. Parameters: a, b
  2. Range （値域）: $[a,b]$
  3. Pdf (density): $f(x) = 1/(b-a)$
  4. Cdf (cumulative distribution function):
     $F(x) = (x-a)/(b-a)$, $a \le x \le b$
  5. Models: all outcomes in the range $[a,b]$ have equal probability. 値域$[a,b]$におけるすべての結果は同じ確率で起こる。

$f(x)$: 0, if $x \notin [a,b]$; If $x \in [a,b]$



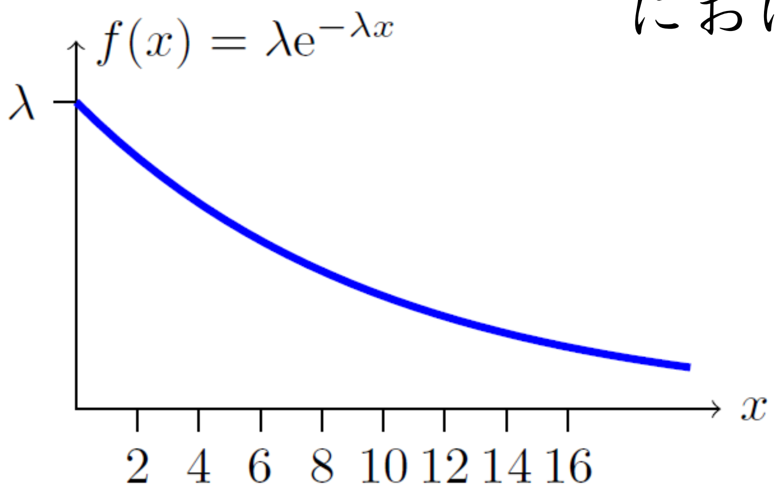
pdf



cdf

Exponential distribution　指数分布

- $X \sim Exponential(\lambda)$ (or $X \sim Exp(\lambda)$)
  1. Parameter $\lambda > 0$
  2. Range $[0, \infty)$
  3. Pdf (density): $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.
  4. Cdf (cumulative distribution function):
     $$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x}$$
  5. Model: waiting time for continuous process (連続過程における待ち時間)



(Exponential distribution) Example I

- Average waiting time in a queue for a taxi at Shinjuku station after the last train is 10min.
  終電後に新宿駅のタクシー乗り場で列に並んで待つ時間は平均的に１０分である。

- Suppose time spent waiting for a taxi is modeled by an exponential random variable
  タクシーを待つ時間は指数分布を用いてモデルされると想定する。

$$X \sim Exp(1/10), \qquad f(x) = \frac{1}{10} e^{-\frac{x}{10}}$$

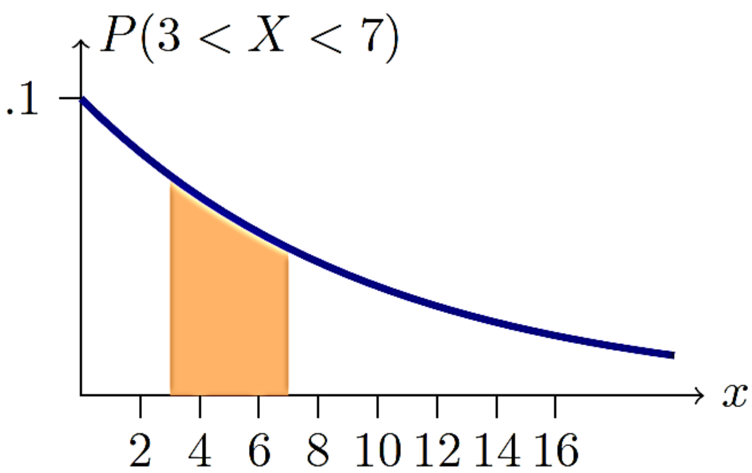
a. Sketch the pdf of this distribution　この分布の確率密度関数を下書きせよ。

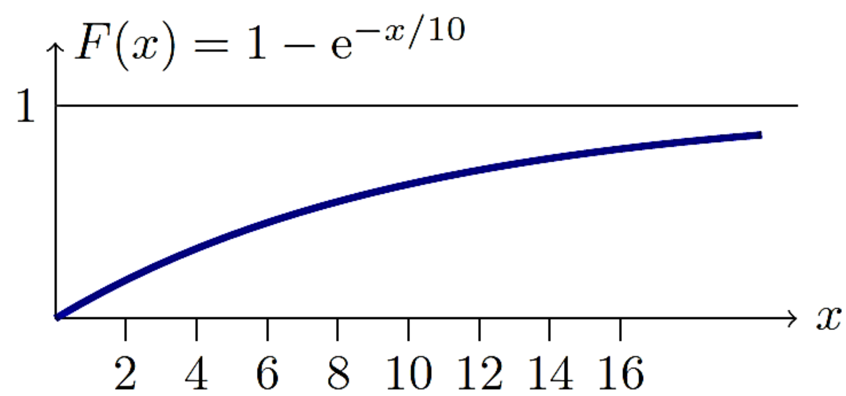
(Exponential distribution) Example II

b. Shade the region which represents the probability of waiting between 3min and 7min.
　３分から7分間の待ち時間を表す領域に色をつけよ。

c. Compute the probability of waiting between 3 and 7min for a taxi
　3分から7分間にタクシーを持つ確率を求めよ。

d. Compute and sketch the cdf
　累積分布を求めて下書きせよ。





Answer (c) : $P(3 < x < 7) = \int_3^7 \frac{1}{10} e^{-\frac{x}{10}} dx$...=.244

Normal (Gaussian) distribution 正規分布

- $X \sim Normal(\mu, \sigma^2)$ (or $X \sim N(\mu, \sigma^2)$)
  1. Parameters: μ, σ
  2. Range 値域: $\mathbb{R} = (-\infty, +\infty)$
  3. Pdf (density): $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
  4. Cdf (distribution): No formula for $F(x)$. We use table or approximations to compute probability.
     $F(x)$ 公式が無い。確率を測定する際に表または近似を利用する。
  5. Models: measurement error, height, averages of lot of data.
     モデル：誤差の測定、身長などの様々なデータの平均

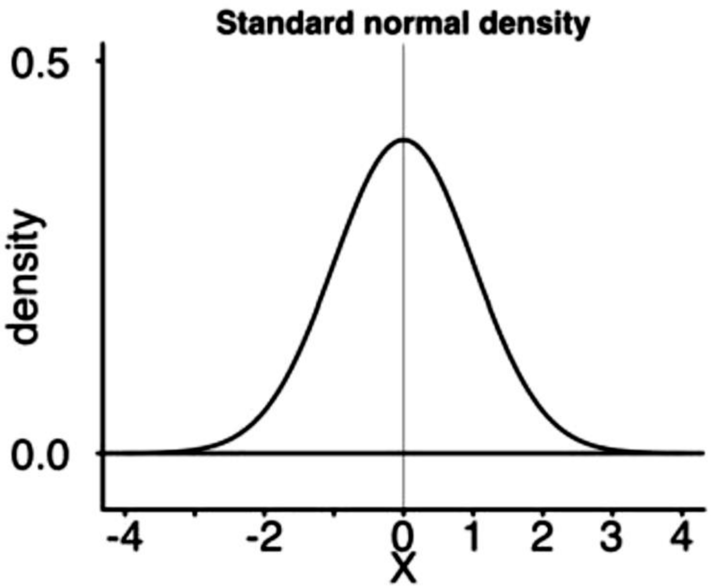
Standard normal distribution 標準正規分布

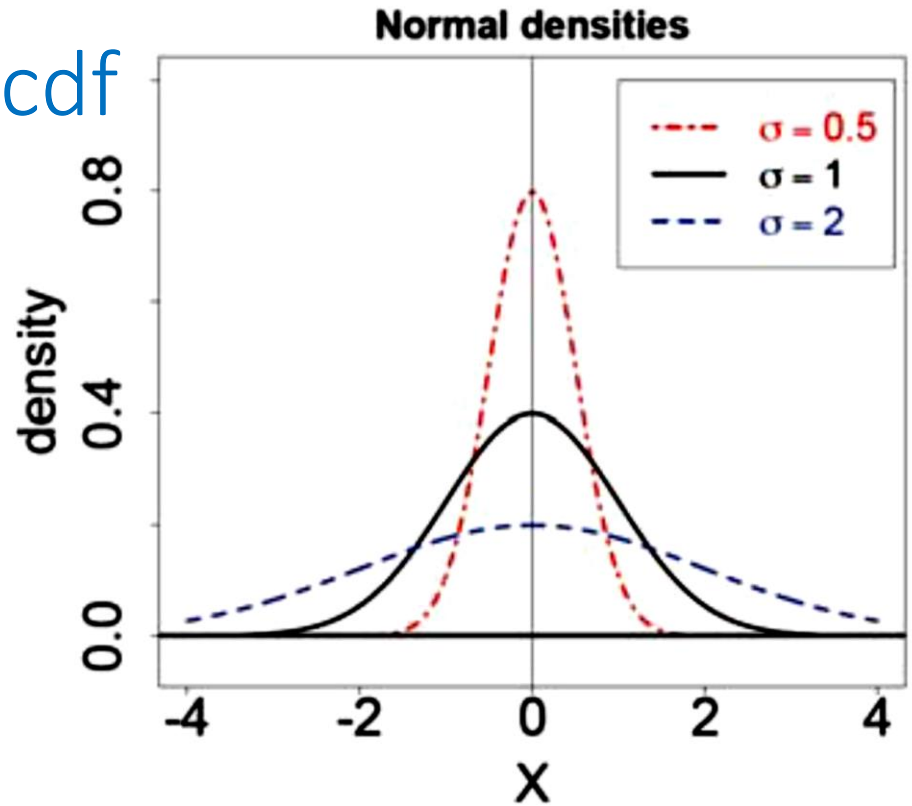
- Standard normal distribution (標準正規分布)
  - Very important special normal distribution commonly denoted Z.
    非常に大事特別な正規分布で、普段にZで記す。
  - $Z \sim N(0,1)$
  - Pdf (probability density function) $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
  - Cdf (cumulative distribution function) commonly denoted $\Phi(x)$ (instead of $F(x)$).
    累積分布は普段に$\Phi(x)$で記す。

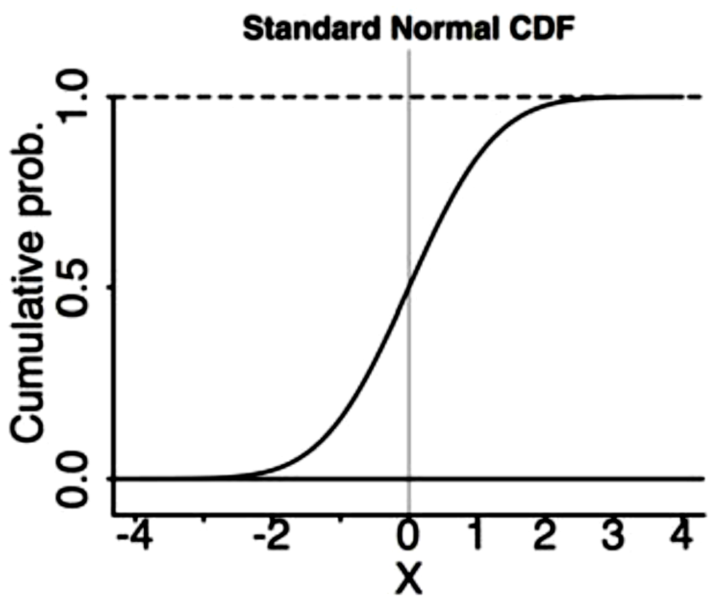
back32

- **Theorem**: If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- Proof: 証明：change of variable (代入積分)

Graphs of pdf and cdf

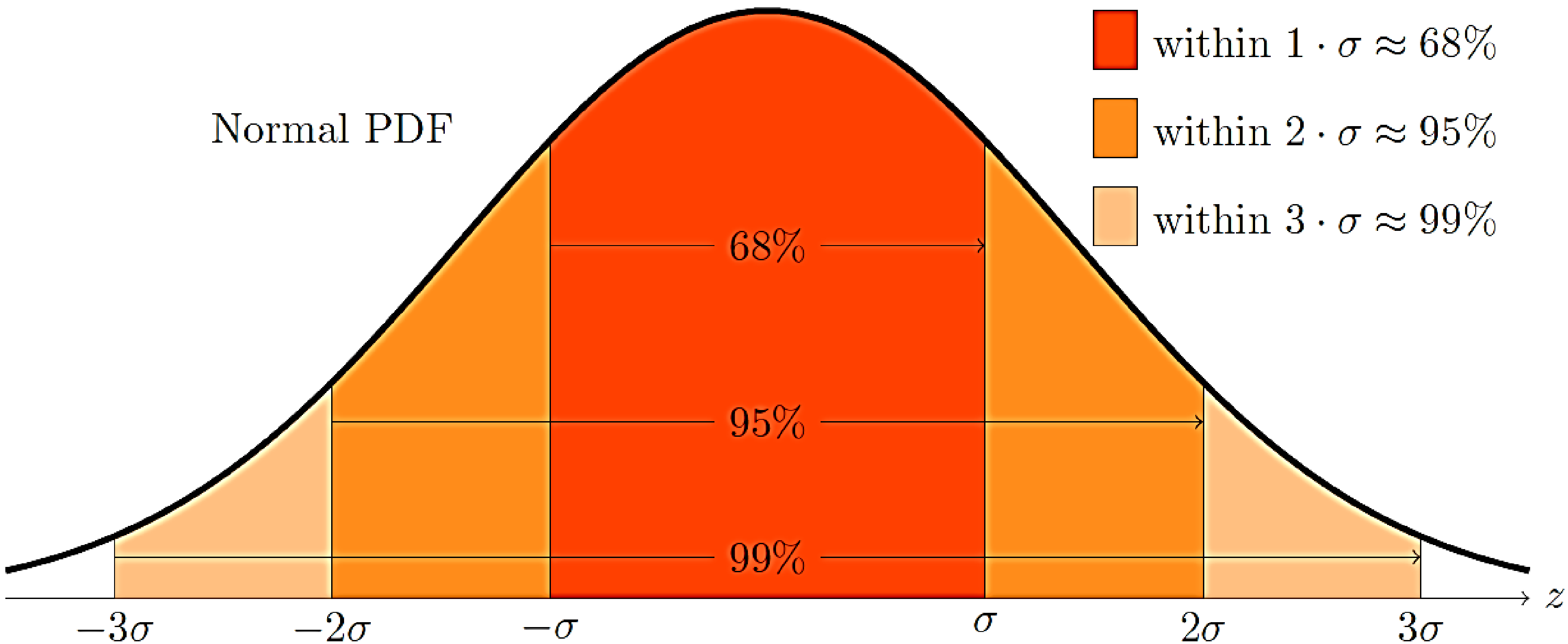






- We'll see later that in a $Normal(\mu, \sigma^2)$ distribution,
  μ is the mean (=expected value)
  σ^2 the variance

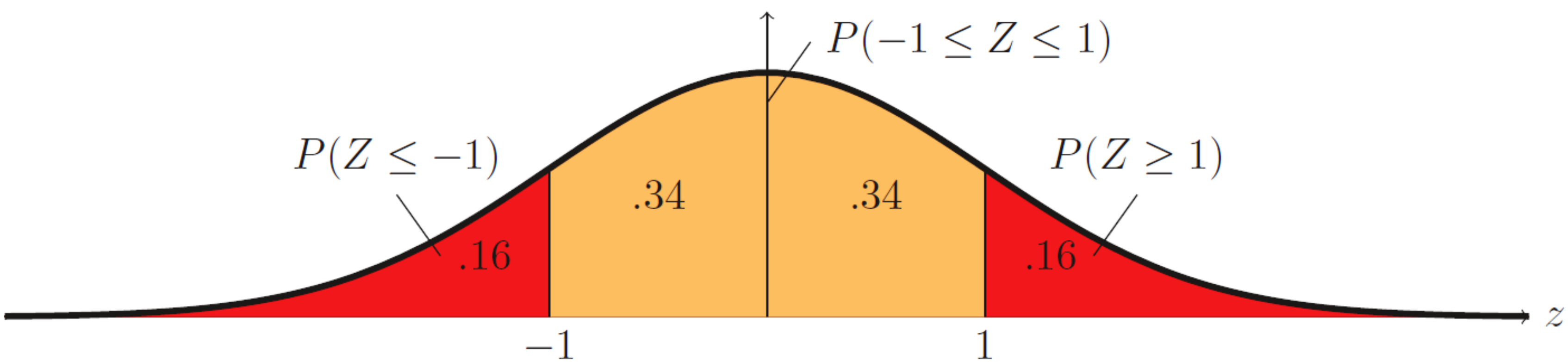
Normal probabilities I



- Rules of thumb （だいたいの目安）　$Z = X - \mu/\sigma$
  - $P(-1 \le Z \le 1) \approx .68$,
  - $P(-2 \le Z \le 2) \approx .95$,
  - $P(-3 \le Z \le 3) \approx .997$

Normal probabilities II

- Use of the symmetry of the density of $N(0, \sigma^2)$.
  $N(0, \sigma^2)$の密度関数の対称性を活かすこと：
  $$f(x) = f(-x)$$
  - **Example:** The rule of thumb says $P(-1 \le Z \le 1) \approx .68$. Use this to estimate $\Phi(1)$ (answer $\Phi(1) \approx 0.84$)



- **Answer:** $\Phi(1) = P(Z \le 1) = P(Z \le 0) + P(0 \le Z \le 1)$
- $P(0 \le Z \le 1) = \frac{1}{2} P(-1 \le Z \le 1) \approx \frac{1}{2} \cdot 0.68 \approx 0.34$
- $\Phi(0) = P(Z \le 0) = 1/2$, $\Phi(1) \approx 0.5 + 0.34 \approx 0.84$

Chapter 2: Random variable

2.8 Expected value and variance (continuous case) 期待値と分散（連続の場合）

- **Expected value**: Measure of location, central tendency 位置の測定、中心傾向
- X continuous random variable with range $[a, b]$ and pdf (= probability density function 確率密度関数) $f(x)$:

$$E(X) = \int_a^b x \cdot f(x)\, dx$$

- Review: If X is discrete with values $x_1, \ldots, x_n$ and pmf (=probability mass function 確率質量関数) $p(x_i)$ then

$$E(X) = \sum_i x_i \cdot p(x_i)\,.$$

- View these two as essentially the same formula. 両方をだいたい同じ式と見なしてもいい。

Variance and standard deviation
分散と標準偏差

- Measure of spread, scale
- For **any** (discrete and continuous) random variable X with mean (=expected value) $\mu = E(X)$,

$$Var(X) = E\left((X-\mu)^2\right), \qquad \sigma = \sqrt{Var(X)}$$

- X continuous with range $[a, b]$ and pdf $f(x)$:

$$Var(X) = \int_a^b (x-\mu)^2 f(x)dx$$

- **Review:** If X is discrete with values $x_1, \dots, x_n$ and pmf $p(x_i)$

$$Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i).$$

- View these two as essentially the same formula.
  両方をだいたい同じ式と見なしてもいい。

Properties of expected value and variance

- Exactly the same as for the discrete case !
  離散の場合とそっくり同じ特性！

1. $E(X+Y) = E(X) + E(Y)$
2. $E(aX+b) = aE(X) + b$ (for numbers a and b)

back32

1. If X and Y are independent, then
   $$Var(X+Y) = Var(X) + Var(Y)$$
2. $Var(aX+b) = a^2 Var(X)$ (for numbers a and b)
3. $Var(X) = E(X^2) - E(X)^2$

Practice Exercise

The random variable X has range [0,1] and pdf $f(x) = cx$.

a. Find c.

b. Find the mean(=expected value), variance and standard deviation of X
X の平均値(=期待値)、分散、標準偏差を求めよ。

c. Suppose $X_1, \ldots, X_{16}$ are independent identically-distributed copies of X. Let $\bar{X}$ be their average. What is the standard deviation of $\bar{X}$?
(定める)分布に従うXと同じ独立同分布に従う確率変数 $X_1, \ldots, X_{16}$ とする。$\bar{X}$ をその平均確率変数とする。$\bar{X}$ の標準偏差はなにか。

d. Suppose Y = X^4. Find the pdf of Y.
(<u>Hint:</u> Start by expressing $P(Y \leq x)$ in function of $P(X \leq x^{1/4}$, then.....)

Solution

Expectation and variance of $U(a,b)$ 一様分布の期待値と分散

- $X \sim Uniform(a,b) \qquad (X \sim U(a,b))$
- $E(X) = \int_a^b x \cdot f(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a}\int_a^b x\,dx$

$$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) = \frac{b+a}{2}$$

- $Var(X) = E(X^2) - E(X)^2 = E(X^2) - \left(\frac{b+a}{2}\right)^2$
- $E(X^2) = \int_a^b x^2 f(x)\,dx = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b$

$$= \frac{1}{b-a}\frac{b^3-a^3}{3} = \frac{b^2+ab+a^2}{3}$$

- $Var(X) = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \cdots = \frac{(b-a)^2}{12}$

Expectated value and variance of $Exp(\lambda)$ 指数分布の期待値と分散

- $X \sim Exponential(\lambda)$ $(X \sim Exp(\lambda))$
- $E(X) = \int_0^\infty x\, f(x)\, dx = \int_0^\infty x\, \lambda e^{-\lambda x} dx$
- Integration by parts (部分積分を用いて解ける)：
- $E(X) = \left[-xe^{-\lambda x}\right]_0^\infty - \int_0^\infty e^{-\lambda x}\, dx = 0 - \left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^\infty = \frac{1}{\lambda}$
- $Var(X) = E(X^2) - \frac{1}{\lambda^2}$
- $E(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx$
- Integration by parts..2 times (2回部分積分) $E(X^2) = 2/\lambda^2$
- $Var(X) = \frac{1}{\lambda^2}$

Expected value and variance of $N(0,1)$ 標準正規分布の期待値と分散

- $X \sim N(0,1)$ $\qquad\qquad \big(X \sim Normal(0,1)\big)$
- $E(X) = \int_{-\infty}^{+\infty} x\, f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-\frac{x^2}{2}}\, dx$
- Integration by substitution (代入法の積分)
- $E(X) = \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{x^2}{2}}\right]_{-\infty}^{+\infty} = 0$

- $Var(X) = \int_{-\infty}^{+\infty} x^2\, f(x)dx - E(X)^2 =$
  $\int_{-\infty}^{+\infty} x^2\, f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}}\, dx$
- Integration by parts : $Var(X) = 1$

Expected value and variance of $N(\mu, \sigma^2)$

- Use Theorem page 20 , the properties of $E(\,.\,)$ and $Var(\,.\,)$ of page 26 , the results of the previous page to show that:
  ページ 20 の定理、 ページ26の $E(.)$と$Var(.)$の性質、前のページの結果を使って以下の命題をみせよ。
  if $X \sim N(\mu, \sigma^2)$ then $E(X) = \mu$, and $Var(X) = \sigma^2$

- *Hint:* Use the random variable $Z = \frac{X-\mu}{\sigma}$.
- $E(Z) =$
- $Var(Z) =$