分布の概要

分布は、ある変数が取り得る値とその頻度を示したものである

重要な記号

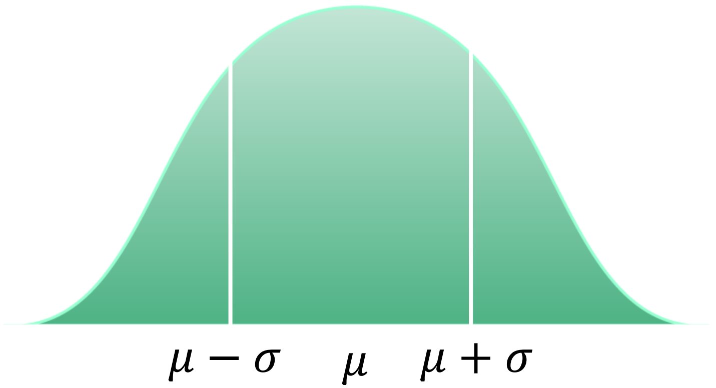
Y 結果

y 結果の内の一つ

$P(Y = y)$ はp(y)と同じ

確率をそれぞれの結果に割り当てる関数のことを確率関数と呼びます

	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s



分布の種類

ある分布は、同じような特徴を有しています。ですので、その分布を種類ごとに分けていきます。また、いくつかの分布は美しい統計量を持っているという特徴があります。まずは、大きなカテゴリとして離散と連続という二種類のカテゴリに分け、それぞれの内容についてみていきます。

離散

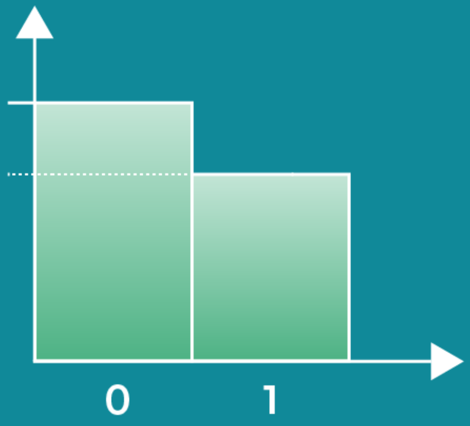
- 有限の結果
- 既に学んできた公式を活用
- ある区間の確率を求める場合は、それらを足していく.
- 表やグラフなどの形で表現をすることができる。
- グラフは棒グラフのような形になる

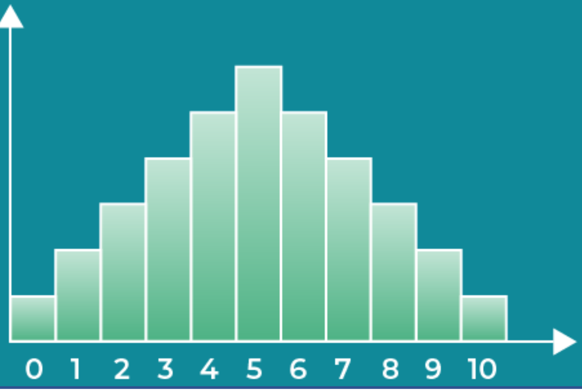
連続

- 連続した無限の結果
- 確率を求めるには、まだ学んでいない公式を使う必要がある
- 数が無限という事から、個別の確率を足すことはできない
- 図や連続した関数で表現することができき
- 図はなめらかな曲線となる

離散分布

離散分布は、求める結果の数が有限の場合を対象とします





離散分布の主な特徴

- 有限の結果
- 既に学んできた公式を活用
- ある区間の確率を求める場合は、それらを足していく.
- 表やグラフなどの形で表現をすることができる。
- グラフは棒グラフのような形になる
- $\boldsymbol{P(Y \leq y) = P(Y < y + 1)}$

離散分布の例:

- 離散一様分布
- ベルヌーイ分布
- 二項分布
- ポアソン分布

一様分布

結果が同様に確からしい場合は一様分布となります

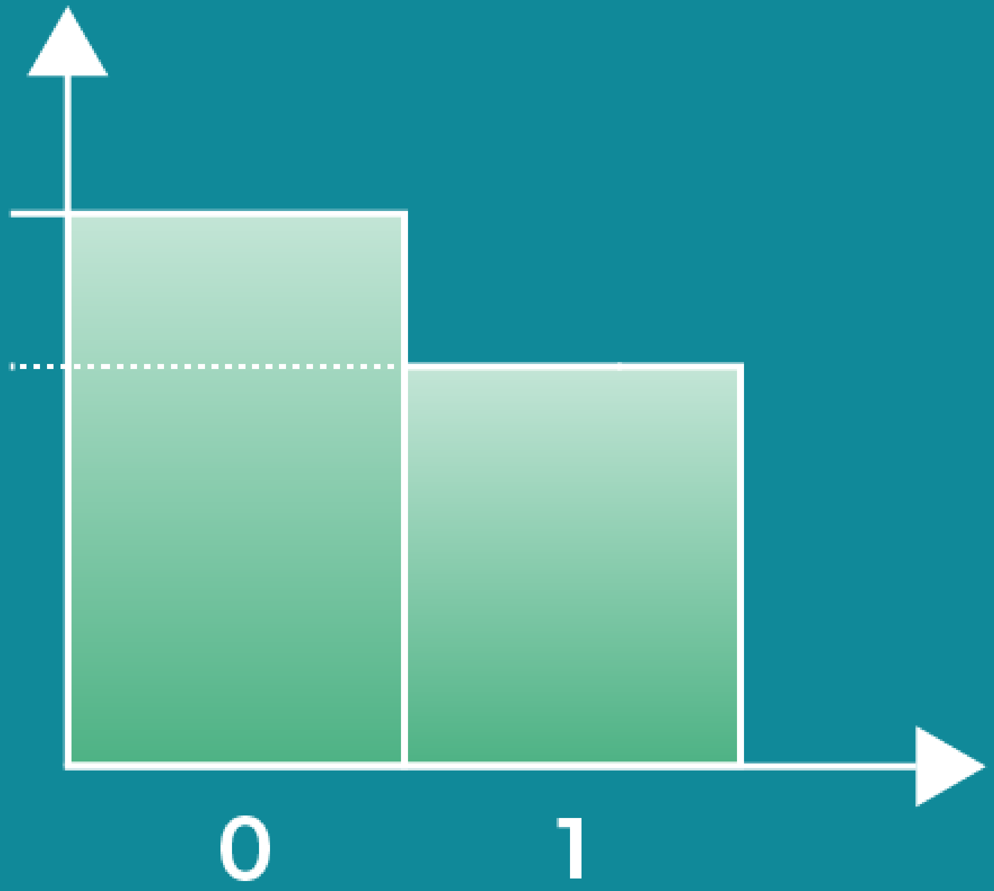
表記法:
- $Y \sim U(a, b)$
- *** 値がカテゴリ変数の場合　$Y \sim U(a)$のようになります**

主な特徴
- 全ての結果は同様に確からしい
- 棒の高さは全て同じ
- 期待値に予測力はない

具体例:
- サイコロを振った結果

ベルヌーイ分布

単一の試行で結果が二つの場合において構成される分布をベルヌーイ分布と言います



表記法:

- $Y \sim Bern(p)$

主な特徴

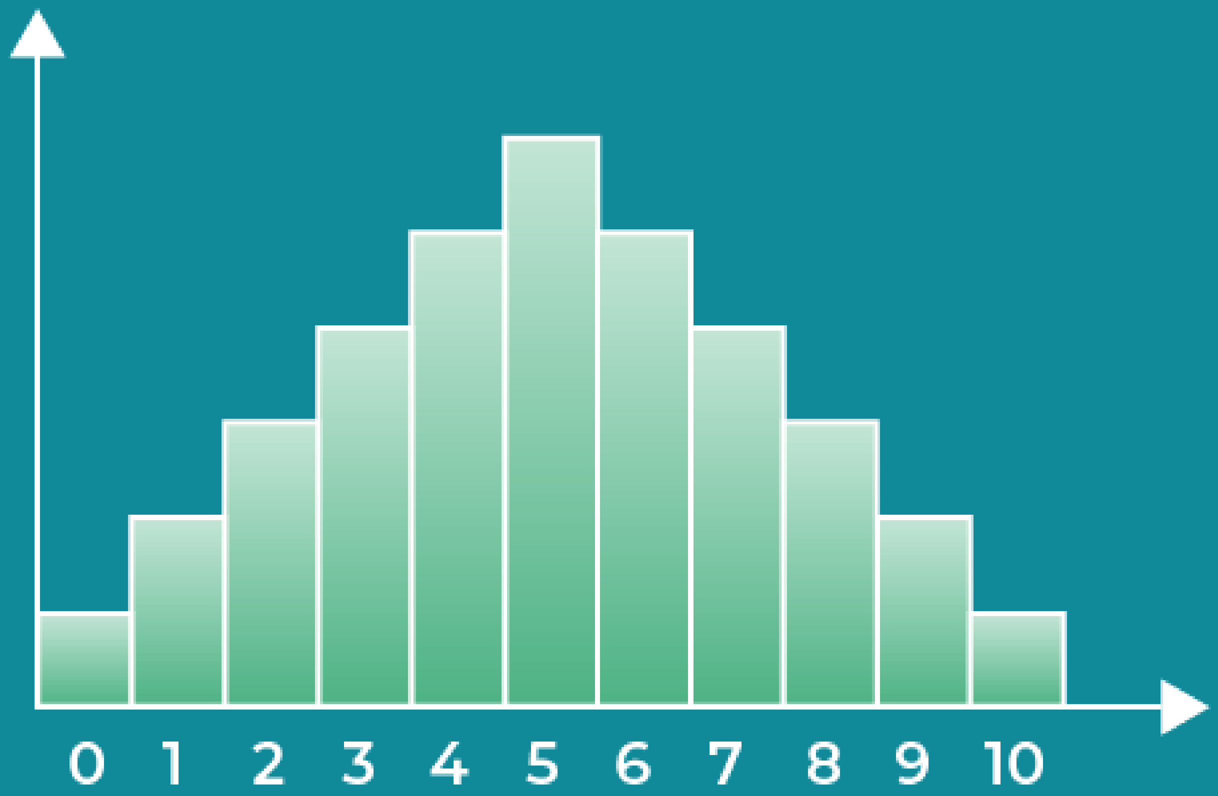
- 試行は一回
- 二つの起こり得る結果
- $E(Y) = p$
- $Var(Y) = p \times (1 - p)$

具体例:

- 二択の質問

二項分布

ベルヌーイ事象を繰り返した結果得られる分布



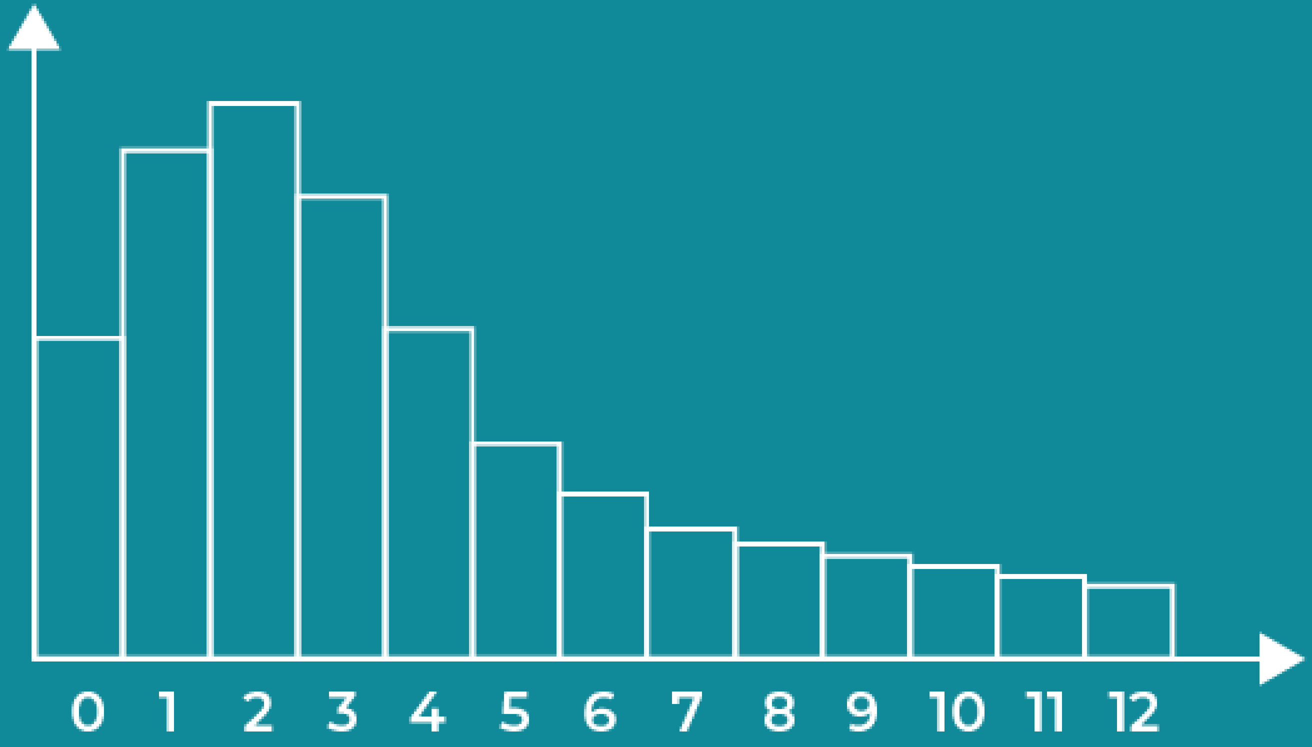
表記法:
- $Y \sim B(n, p)$

主な特徴
- n 回試行をした場合における確率を求めるもの
- $P(Y = y) = C(y, n) \times p^y \times (1 - p)^{n-y}$
- $E(Y) = n \times p$
- $Var(Y) = n \times p \times (1 - p)$

具体例:
- コインを10回投げた場合に何回表が出るか

ポアソン分布

ある期間において事象が起こる確率を示した分布をポアソン分布という



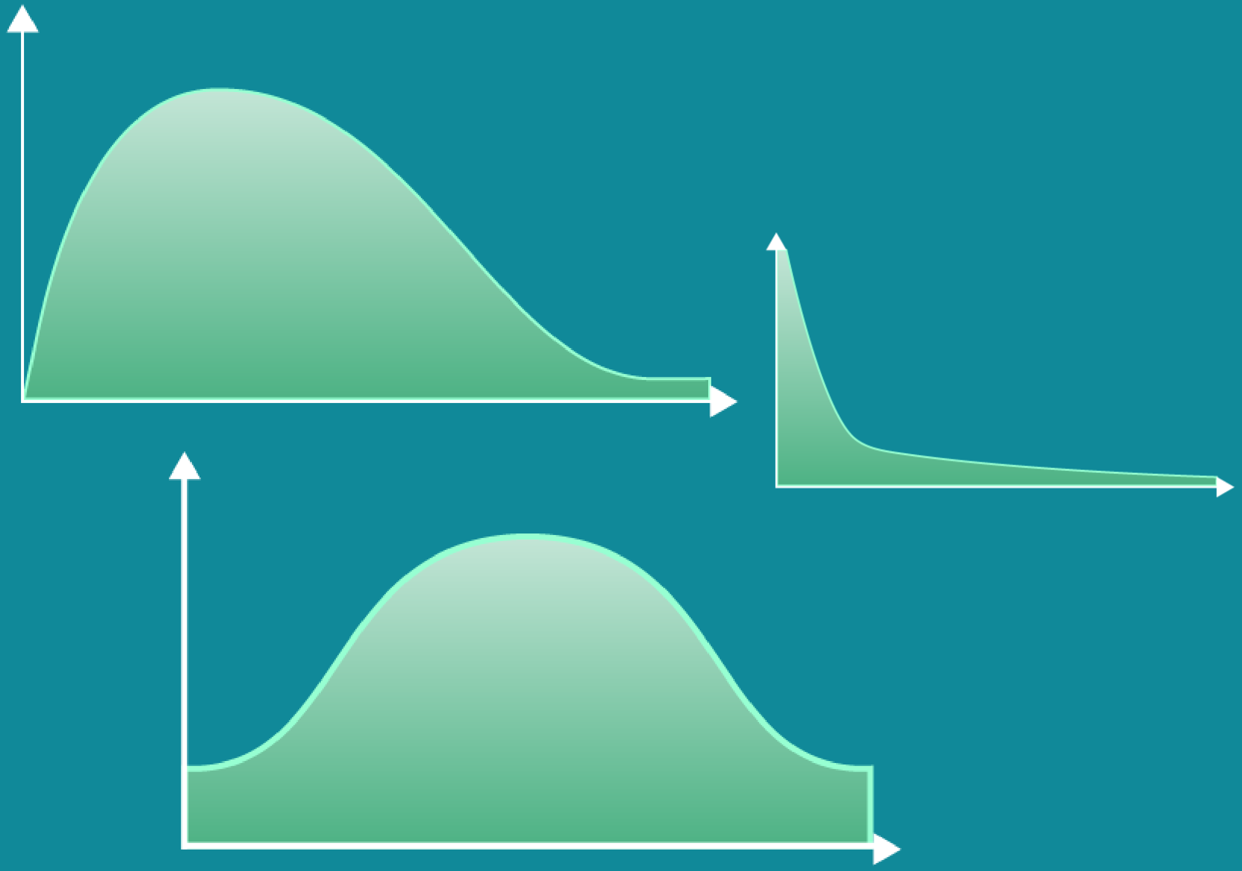
表記法:
- $Y \sim Po(\lambda)$

主な特徴
- ある期間において事象が発生する確率を示す
- $P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$
- $E(Y) = \lambda$
- $Var(Y) = \lambda$

具体例:
- 10秒の間にホタルが光る確率

連続分布

取り得る値が連続である場合、連続分布を用いる

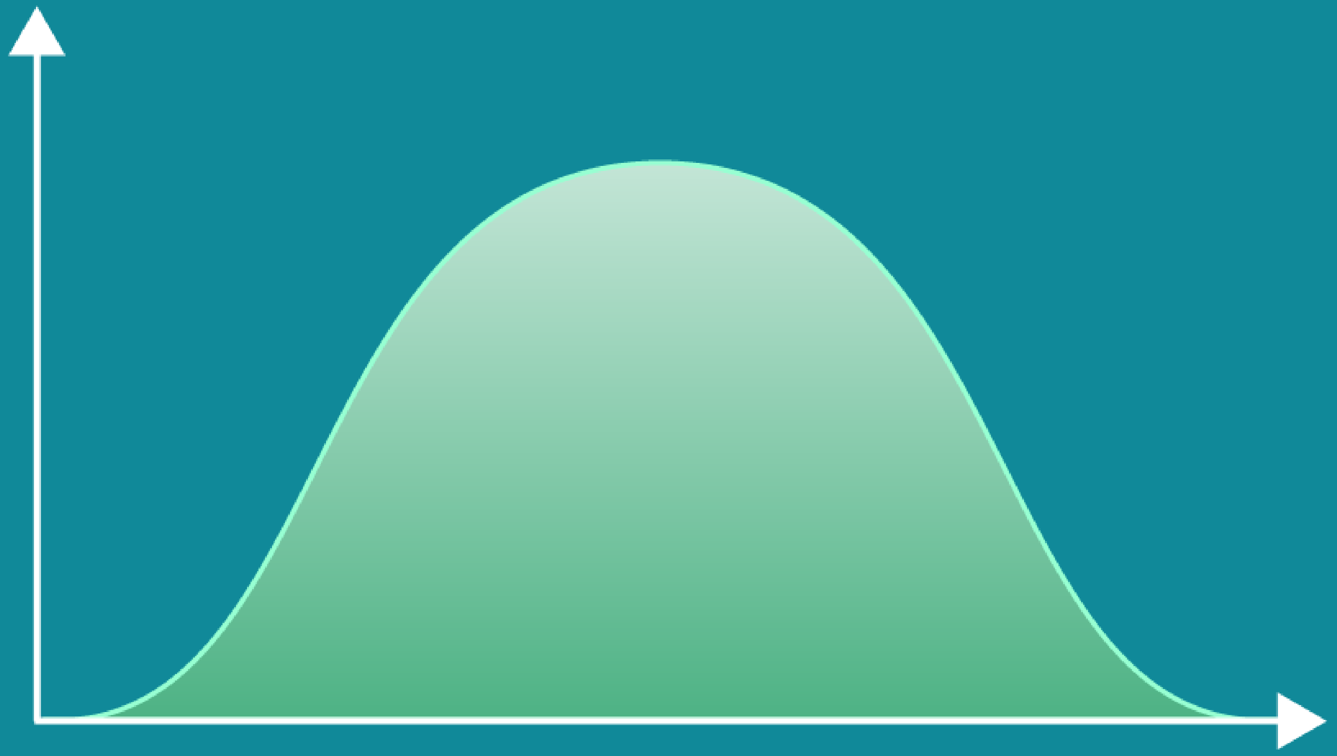


Key characteristics

- 連続した無限の結果
- 確率を求めるには、まだ学んでいない公式を使う必要がある
- 数が有限という事から、個別の確率を足すことはできない
- 図や連続した関数で表現することができき
- 図はなめらかな曲線となる
- 確率を計算するには積分を用いる必要がある
- 累積分布関数（CDF）を有する
- $P(Y = y) = 0$
- $P(Y < y) = P(Y \leq y)$

正規分布

正規分布は実生活において良く見られる分布となる



表記法:

- $Y \sim N(\mu, \sigma^2)$

主な特徴

- グラフはベルの形をしており、左右対称で両端が薄い
- $E(Y) = \mu$
- $Var(Y) = \sigma^2$
- 68% のデータが以下の範囲に入る
  - $(\mu - \sigma,\ \mu + \sigma)$

具体例:

- 自然界の動物の体重で見られることがある

正規分布の標準化

正規分布を標準化することによって、平均が0、標準偏差が0の分布を得ることができる

変換によって作成された新しい変数z

$$z = \frac{y - \mu}{\sigma}$$

平均

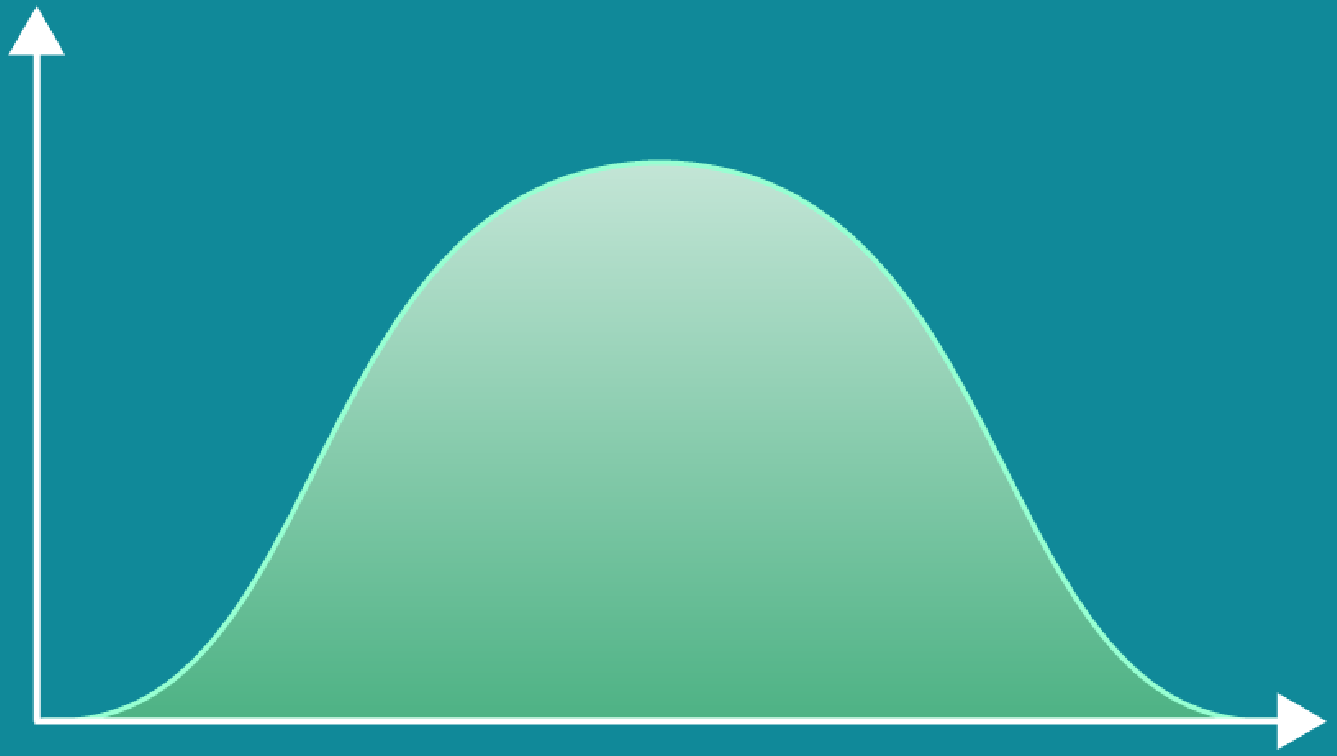
標準偏差

標準正規分布の重要性

- 標準正規分布を用いることで、それぞれの値が平均からどれぐらい離れているかを知ることができる
- 上記の公式を使うことで、いかなる正規分布も標準正規分布にすることができる
- そのCDFはZスコア表と呼ばれている（もしくはZ表）

スチューデントのT分布

標本が少ない場合に正規分布を模した分布



表記法:

- $\boldsymbol{Y \sim t\ (k)}$

主な特徴

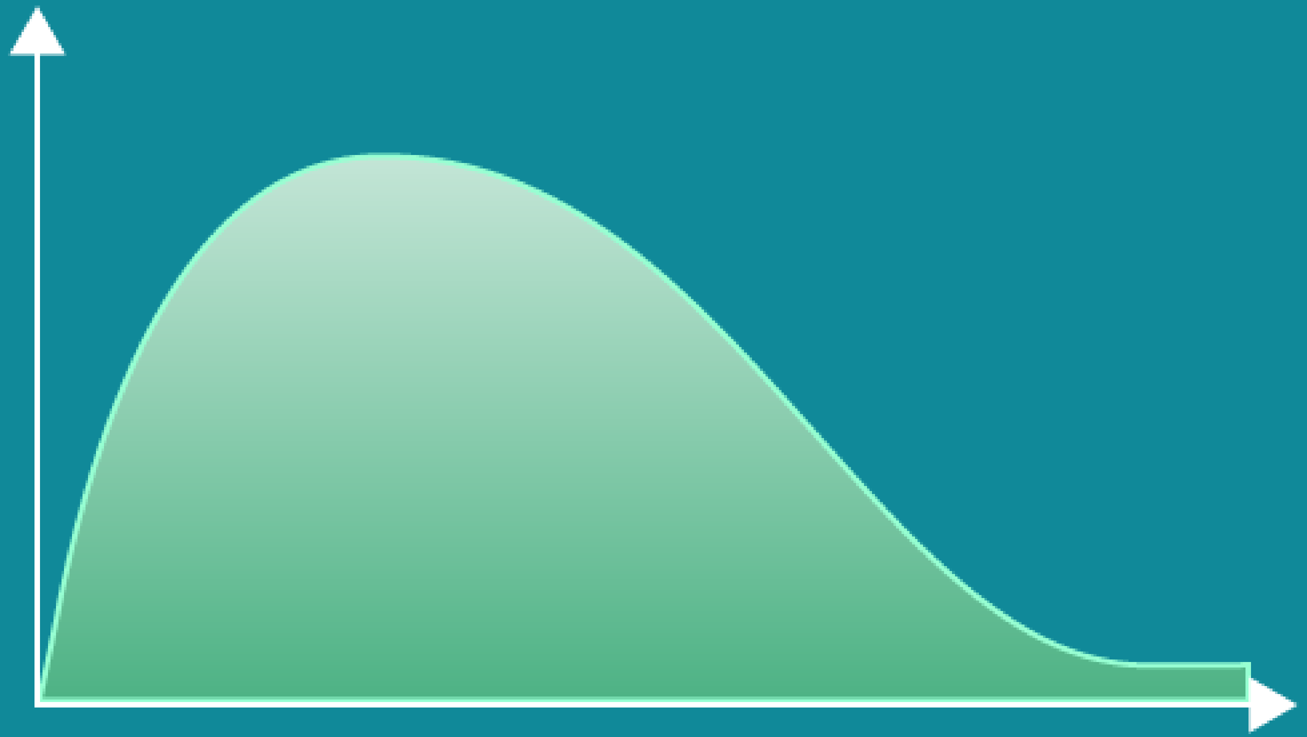
- 少ない標本で正規分布を近似する
- グラフはベルの形をしており、左右対称であるが、両端が厚い
- 両端の確率がより高いとする
- If k>2: $\boldsymbol{E(Y)} = \mu$ and $\boldsymbol{Var(Y) = s^2 \times \frac{k}{k-2}}$

具体例:

- 標本の数が少なく、正規分布を想定することができる場合に使われる

カイ二乗分布

あまり使われない分布



表記法:

- $Y \sim \chi^2(k)$

主な特徴

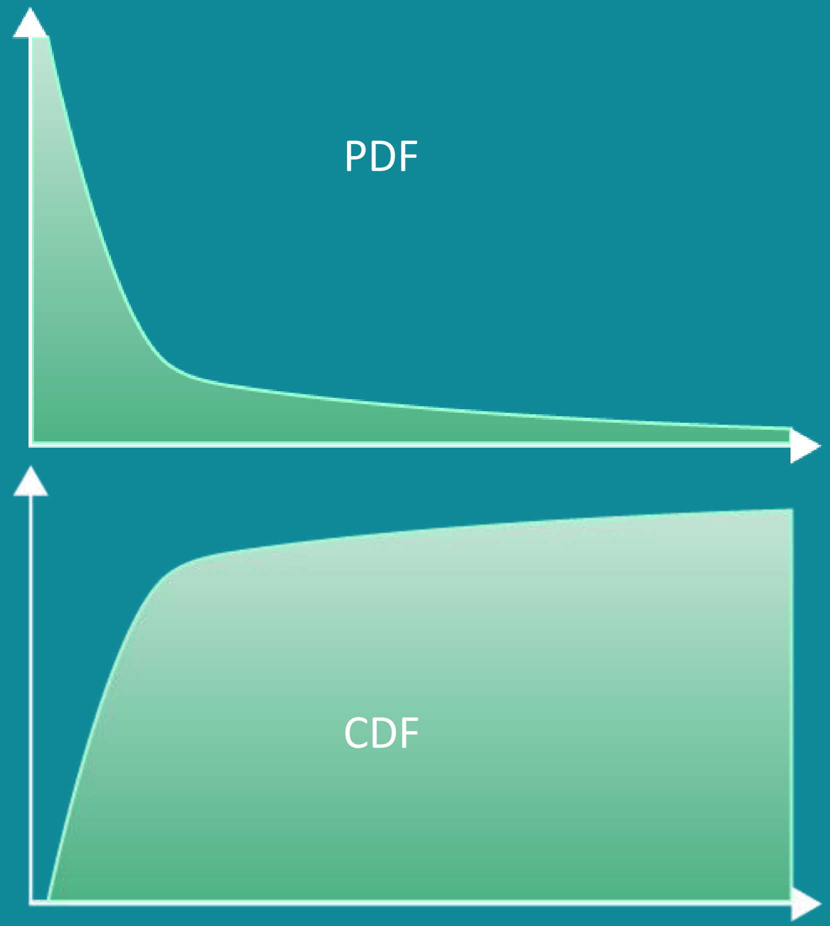
- グラフは非対称であり、右に歪んでいる
- $E(Y) = k$
- $Var(Y) = 2k$
- t 分布の累乗の分布である.

具体例:

- 当てはまりの良さを調べる場合に使われる
- CDFはχ^2表と呼ばれている

指数分布

数字が急速に変わる場合に見受けられる分布



表記法:

- $Y \sim Exp\ (\lambda)$

主な特徴

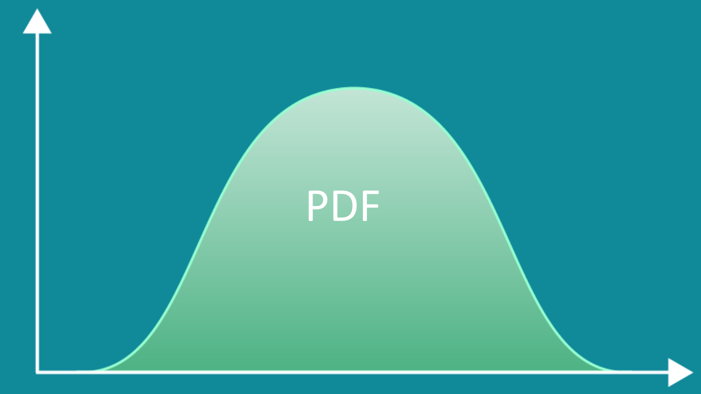
- PDF（確率密度関数）とCDFはある値ではほとんど変わらなくなる
- $E(Y) = \frac{1}{\lambda}$
- $Var(Y) = \frac{1}{\lambda^2}$
- カイ二乗などの数字がまとまった表がないため、対数変換などを行う場合がある

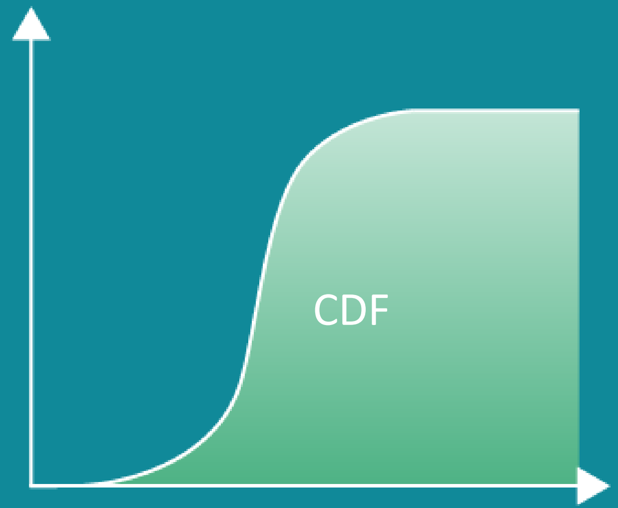
具体例:

- ウェブサイトのPVなど、値が急速に変化する場合に使われる

ロジスティック分布

二択の判別をする場合の判別などにおいて使われる分布





表記法:

- $Y \sim Logistic\ (\mu, s)$

主な特徴

- $E(Y) = \mu$
- $Var(Y) = \frac{s^2 \times \pi^2}{3}$
- CDFは、平均近くで値が一気に上がる
- スケールパラメーターが小さいほど、1にすぐに近づいていく

具体例:

- スポーツであるチームや選手の勝敗について整理する場合などに使われる