2024年度春学期

# 統計学

第11回

分布の「型」を考える

— 確率分布モデルと正規分布


関西大学総合情報学部
浅野 晃


---

## ちょっと前回の復習

---

## 「統計的推測」とは

調べたいデータ全体を調べられるか？

日本男性全員の身長を調べられるか？

データの一部を調べて度数分布を推測する

いや, せめて平均や分散を推測する

統計的推測



---

## 無作為抽出

データ全体から, いくつかの数値を
公平なくじびきで選ぶ

[無作為標本抽出]という

調べたい(が全部を調べるのは無理な)集団[母集団]

調べられる程度のデータ[標本(サンプル)]

## 度数分布で考えると

| 母集団の度数分布 | |
|---|---|
| 階級値 | 相対度数 |
| 162.5 | 15% |
| 167.5 | 20% |
| 172.5 | 20% |
| 177.5 | 10% |

$\xrightarrow{\text{無作為抽出}}$

| 標本の[確率分布] | |
|---|---|
| 階級値 | 選ばれる確率 |
| 162.5 | 15% |
| 167.5 | 20% |
| 172.5 | 20% |
| 177.5 | 10% |



## 確率分布と確率変数

つまり 母集団の度数分布（母集団分布） ＝ 標本の確率分布

| 階級値 | 選ばれる確率 |
|---|---|
| 162.5 | 15% |
| 167.5 | 20% |
| 172.5 | 20% |
| 177.5 | 10% |

標本は，
値がいくらになるかは決まっていない
しかし確率分布が決まっている
（知っているかどうかは別）

こういう数を[確率変数]という
（中国語では随機変数）

「標本は，確率変数（の一種）である」



## 母平均の推定

母集団
（日本男性全体）

母平均 $\mu$

標本として数値をいくつか取り出して，
それらの平均 ［標本平均］

標本平均は母平均に近い値になるか？

母平均が知りたい が，日本男性全員は調べられない



## 標本平均の期待値と分散は

母集団

母平均 $\mu$
母分散 $\sigma^2$

サイズ $n$ の標本1セット　標本平均

$X_1 \quad X_2 \quad \cdots \quad X_n \qquad \bar{X}$

$X_1 \quad X_2 \quad \cdots \quad X_n \qquad \bar{X}$

$X_1 \quad X_2 \quad \cdots \quad X_n \qquad \bar{X}$

$\vdots \qquad\qquad\qquad\qquad \vdots$

母集団と同じ
期待値 $\mu$
分散 $\sigma^2$

極端な値はあまりないので
分散が小さくなる

期待値 $\mu$
分散 $\sigma^2/n$

標本平均の分散は，母分散の「標本サイズ分の一」になる

## 母平均の推定

母平均が $\mu$
母分散が $\sigma^2$ のとき，
標本平均の期待値が $\mu$
標本平均の分散が $\sigma^2/n$

仮に，何度も標本を抽出して，何度も標本平均を計算したとすると

分散が小さくなっているので，「たいてい，ほぼ」母平均に近い

? ? ? ?
$\mu$ —— $\bar{X}$ $\bar{X}$ $\bar{X}$ $\bar{X}$ ——→

標本平均を
何度も計算しても

いつ計算しても，たいていそれほど変わらない

いま1回だけ計算した標本平均は，上のどれなのかわからないが
たいてい，ほぼ母平均に近い値だろう



## 「標本の大きさ」の意味

母分散が $\sigma^2$ のとき，標本平均の分散が $\sigma^2/n$

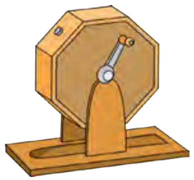
標本平均の分散に関係しているのは
標本の大きさであって，母集団の大きさは関係ない

推測の確かさに影響するのは
標本の大きさであって，
標本の大きさの，母集団の大きさに対する割合 ではない



## 標本の大きさとは

「10人からなる標本」の意味は，
1,000人からなる母集団でも100,000人からなる母集団でも同じ
🤔…



理想的な無作為抽出では，復元抽出を行う
標本サイズは，
「取り出された数値の個数」というよりも
「同一の母集団から数値ひとつひとつを取り出す回数」
→ 母集団の大きさに対する割合は無関係

（非復元抽出をした場合は，計算で補正する方法がある）



## 母平均の推定

いま1回だけ計算した標本平均は，
「たいてい，ほぼ」母平均に近い値だろう

どのくらい近い？

どのくらいの確率で？
はずれる確率は？

ここから先に進みます。

分布の「型」を考える🤔

## 母平均の推定

いま1回だけ計算した標本平均は，
「たいてい，ほぼ」母平均に近い値だろう

どのくらい近い？

どのくらいの確率で？
はずれる確率は？

計算するには，
式で表されてないといけない



## 確率分布と確率変数

つまり 母集団の度数分布（母集団分布）＝ 標本の確率分布

| 階級値 | 選ばれる確率 |
|---|---|
| 162.5 | 15% |
| 167.5 | 20% |
| 172.5 | 20% |
| 177.5 | 10% |

これは式ではなく数値の集まり，
計算できない

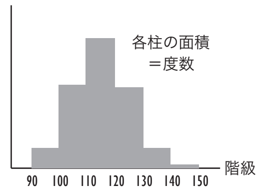


## 式で表す

度数分布を

| 階級値 | 選ばれる確率 |
|---|---|
| 162.5 | 15% |
| 167.5 | 20% |
| 172.5 | 20% |
| 177.5 | 10% |

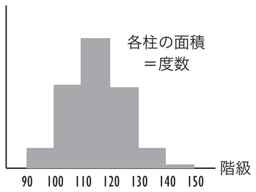
ヒストグラムが




何かの式で書ける
ものと仮定する

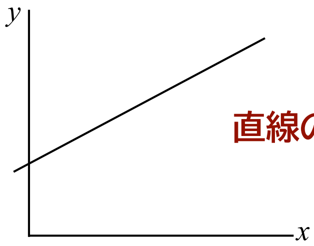
何かの式で表される関数の
グラフであると仮定する

## 確率分布モデルとパラメータ




何かの式のグラフで
あると仮定する

式＝［確率分布モデル］




直線のモデル

$$y = ax + b$$

パラメータ

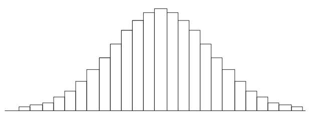
パラメータを推定すればグラフが描ける

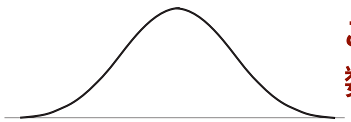


## 連続型確率分布📈

## ヒストグラムを式で表す

こんなヒストグラムを，
式で書けるだろうか？





これを表す式のほうが
数学は簡単。

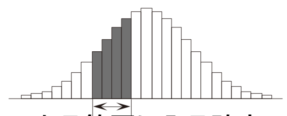
階級の区切り方が
どんどん細かくなって，
見えなくなったと考える

［連続型確率分布］



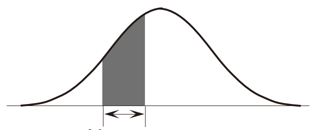
## 連続型確率分布

ヒストグラム



ある範囲に入る確率
＝柱の面積の合計

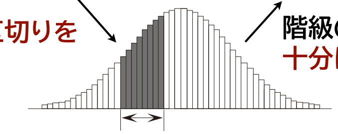
［連続型確率分布］



同じ範囲なら
柱の面積の合計は同じ

階級の区切りを
細かく

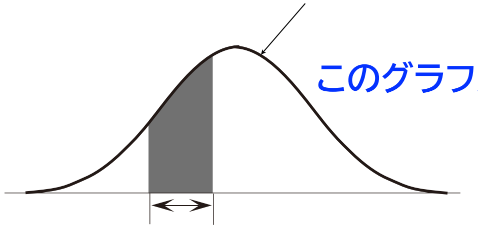
階級の区切りを
十分に細かく



同じ範囲なら
柱の面積の合計は同じ

## 確率密度関数と確率

ヒストグラムの上の縁＝[確率密度関数]

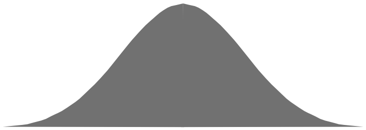


このグラフが示すのは確率ではない

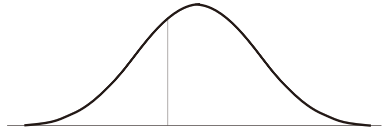
この範囲に入る確率＝この面積
＝確率密度関数の積分



## 確率密度関数の矛盾？



連続型確率変数が
すべての実数のうちのどれかになる確率
＝1（100%）



$a$

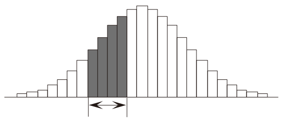
幅が0だから，面積も0

連続型確率変数が
ある特定の値 $a$ になる確率
＝0

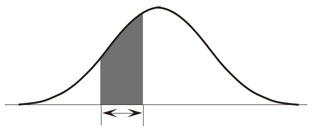
なんかヘン？付録テキストで



## 連続型確率分布は，数学の都合



こんなのより



こんなののほうが数式にしやすい

実際のデータは，有限の桁数の数字で
表されている限り，必ず離散的。



# 正規分布モデル

# 正規分布モデル

世の中には，[正規分布モデル]で表せるような母集団分布がたくさんある

長さの測定値の分布，共通テストの成績の分布　…

[中心極限定理]

母集団のばらつきの原因が，無数の独立な原因の和のとき，母集団分布は概ね正規分布になる

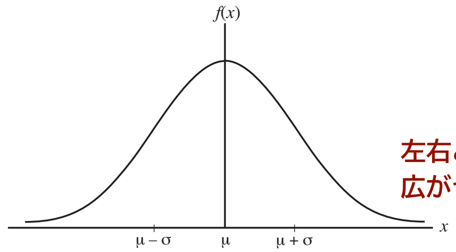
「日本男性の身長」も。



# 正規分布の特徴

パラメータが$\underbrace{\text{平均(期待値)}}_{\mu}$と$\underbrace{\text{分散}}_{\sigma^2}$

確率密度関数はこんな形




左右とも無限に広がっている



# 正規分布の特徴

パラメータが$\underbrace{\text{平均(期待値)}}_{\mu}$と$\underbrace{\text{分散}}_{\sigma^2}$

（わかりやすいものを推定すればよいので都合がいい）

確率変数 $X$ の確率分布が
期待値 $\mu$，分散 $\sigma^2$ の正規分布であることを

確率変数 $X$ が $N(\mu, \sigma^2)$ にしたがう　という

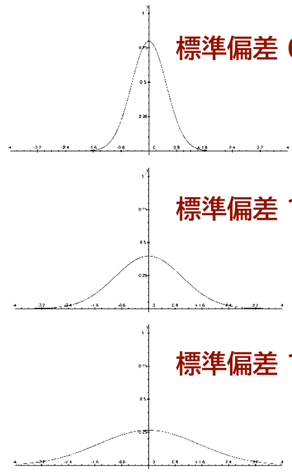
※英語ではnormal distribution，中国語では「常態分配」



# 正規分布の特徴

期待値0の正規分布の確率密度関数

標準偏差が大きくなると
中央部の広がりが大きくなり
高さが低くなる



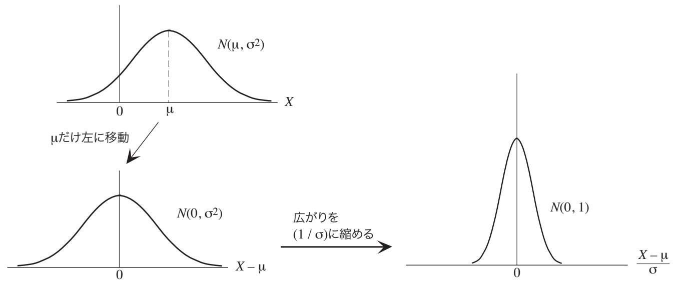
標準偏差 0.5

標準偏差 1.0

標準偏差 1.5

## 正規分布の性質1

確率変数 $X$ が $N(\mu, \sigma^2)$ にしたがう　とき




$(X-\mu)/\sigma$ は $N(0,1)$ にしたがう



## 正規分布の性質1

確率変数 $X$ が $N(\mu, \sigma^2)$ にしたがう　とき

$(X-\mu)/\sigma$ は $N(0,1)$ にしたがう

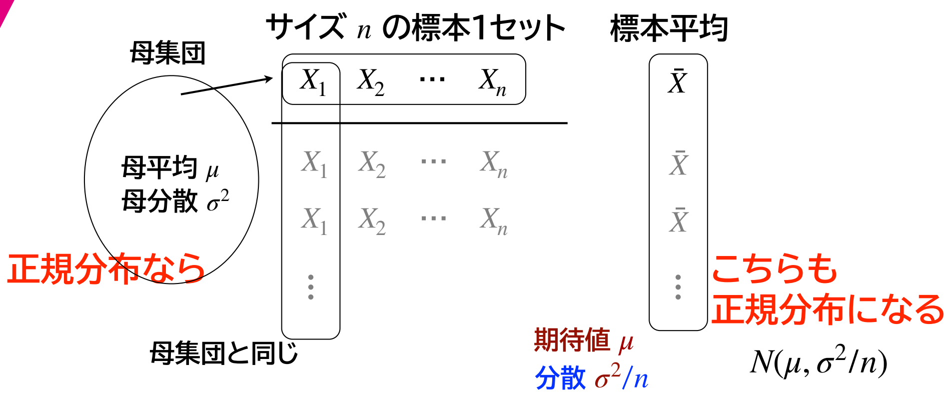
「標準得点」と同じ

変換しても，
やはり正規分布になる

$N(0,1)$ を[標準正規分布]という



## 正規分布の性質2






## 正規分布表の使いかた🔢

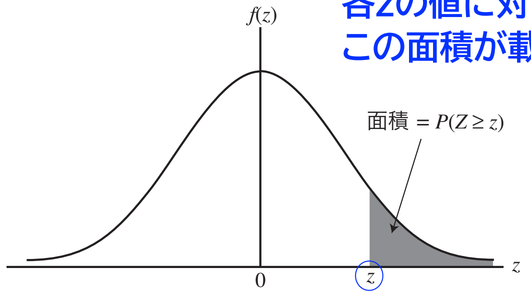
「数表」を使って手作業で計算することは，いまではなくなりましたが，
その仕組みだけは知っておくとよいと思います。

## 正規分布にもとづく計算

正規分布にしたがう確率変数がある範囲に入る確率

数表を使って求める

標準正規分布について,
各zの値に対する
この面積が載っている






## 正規分布にもとづく計算

例)確率変数 $X$ が $N(50,10^2)$ にしたがうとき,
$X$ が 60 以上である確率を求めよ。

性質1により, $Z = (X - 50)/10$ と変換
$Z$ は標準正規分布にしたがう

$X = 60$ のとき, $Z = (60 - 50)/10 = 1$

よって, 求めるのは, $Z$ が 1 以上である確率 $P(Z \geqq 1)$



## 正規分布にもとづく計算

$P(Z \geqq z)$ を求める

$z$ の小数第2位

$z$ の小数第1位まで

| | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | |
|---|---|---|---|---|---|---|---|
| 0.0 | 0.50000 | 0.49601 | 0.49202 | 0.48803 | 0.48405 | 0.48006 | |
| 0.1 | 0.46017 | 0.45620 | 0.45224 | 0.44828 | 0.44433 | 0.44038 | |
| 0.2 | 0.42074 | 0.41683 | 0.41294 | 0.40905 | 0.40517 | 0.40129 | ··· |
| 0.3 | 0.38209 | 0.37828 | 0.37448 | 0.37070 | 0.36693 | 0.36317 | |
| 0.4 | 0.34458 | 0.34090 | 0.33724 | 0.33360 | 0.32997 | 0.32636 | |
| | | | ⋮ | | | | |
| 1.0 | 0.15866 | 0.15625 | 0.15386 | 0.15151 | 0.14917 | 0.14686 | ··· |

$P(Z \geqq 1)$



## 正規分布にもとづく計算

テキストの「問題例」

確率変数 $X$ が正規分布 $N(50,10^2)$ に
したがうとき, 次の確率を求めてください。

(1) $P(X \geqq 55)$ (2) $P(45 \leqq X \leqq 60)$

$X$ が 55 以上である確率　$X$ が 45 以上で 60以下である確率

考え方

$Z = (X - 50)/10$ と変換すると, 正規分布の性質1から,
$Z$ は標準正規分布 $N(0,1)$ にしたがう

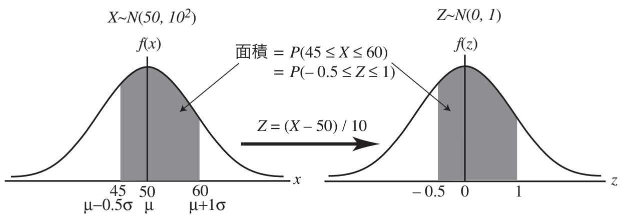
# 正規分布にもとづく計算

「問題例」の2

$X = 45$ のとき　$Z = (45 - 50)/10 = -0.5$

$X = 60$ のとき　$Z = (60 - 50)/10 = 1$

よって $P(45 \leqq X \leqq 60) = P(-0.5 \leqq Z \leqq 1)$




このグレーの部分の面積をどうやって求める？



# 正規分布にもとづく計算

「問題例」の2　パズルをおこなう