第 2 章　バラついた数を読む
ー分布の話ー

1. パン屋のインチキを暴く話

米をパンに切り替えたら血圧が下がるのか？

白米に比べライ麦パンなどはカリウムを多く含むため、血圧上昇の予防になる。パンでなくとも玄米でも可。

ヒストグラム

「横軸にパンの重さを取り、縦軸にパンの数をとって (p38)」描いた次のようなグラフをヒストグラムと呼ぶ。

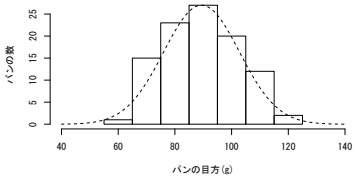

図 1. ヒストグラム

本文中では 60,70,80... グラムのパンの重さを数えたと書いてあるが、実際にはパンの重さがそのような丁度の値だけをとるということはありえない。本文中表 10(p39) は正しくは次のような形となる。

表 1. 度数分布表

階級 (g)	階級値 (パンの重さ,g)	度数 (個数, 個)
55 - 65	60	1
65 - 75	70	15
75 - 85	80	23
85 - 95	90	27
95 -105	100	20
105 -115	110	12
115 -125	120	2

このように、ある代表値 (階級値) を中心とするある範囲 (階級) に含まれるデータの個数 (度数) をまとめた表を度数分布表と呼ぶ。

図 1 に点線で示したのはパンの数を無限大、階級幅を無限小にしたときのヒストグラムで、この曲線により示される分布を正規分布、Gauss 分布などと呼ぶ。

データの切捨てのような人為操作が加わると、本文中表 12(p40) のように切断された分布となる。これは切断正規分布と呼ばれる。

コントロールできない偶然の誤差は結果として正規分布を生じやすい。これを逆に利用し、正規性を調べることで偶然でない誤差を探すこともできる。

試料と母集団

「パン屋が売っている全体のパン (p44)」のように、全ての対象の集まりを母集団と呼ぶ。そして、「A 老人が買った 100 切れのパン」のように母集団から抽出された集まりを試料、または標本と呼ぶ。

母集団の概念を使って仮説検定を説明すると、仮説検定というのは「ある試料がある母集団から抽出されたものとみなすことができるかどうか」を調べているということができる。

分布の代表値

平均値は分布がどこを中心としているかを述べるのに便利なパラメータであるが、同じ平均値でも分布が同じとは限らない。

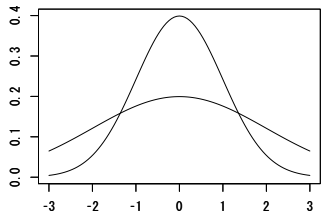

図 2. 2 つの正規分布

図 2 はいずれも平均値が等しい正規分布であるが、明らかに形が違う。とがった分布は平均に近い値が出やすく、平らな分布は出にくい。バラつかない分布とバラついた分布と言い換えることもできる。「バラつき」を測る測度が必要である。そこで、次

の式で定義される母分散 σ^2 を導入する[*1] 。

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \tag{1}$$

n は母集団を構成するデータの数、x_i は個々のデータ、m は母平均値である。母分散は「母平均から個々のデータまでの距離の二乗」の平均値である。二乗ではイメージがしにくいので、それの平方根を取って標準偏差 σ を定義する。

$$\sigma = \sqrt{\sigma^2} \tag{2}$$

正規分布の形は平均値 m と標準偏差 σ(または母分散) の 2 つの情報があれば完全に決定される。

正規分布の基準化

正規分布曲線を計算によって求めるのは多少骨が折れる。そのため、正規分布するデータに適当な変換を施して、何か 1 つの基準となる正規分布に変換できれば便利である。

まずは平均値を 0 にする。これには、個々のデータ全てから母平均を引けばいい。

$$X_i = x_i - m \tag{3}$$

こうして作られた X_i は平均が 0 の分布となる。

次に、標準偏差を 1 にする。これは、それぞれの X_i を標準偏差 σ で割ればいい。

$$u = \frac{X_i}{\sigma} \tag{4}$$

この作業を施せば、どんな分布でも (正規分布でありさえすれば) 平均が 0、標準偏差が 1 の正規分布となる。この正規分布を基準正規分布、もしくは標準正規分布と呼ぶ。そして、今と逆の操作を施せば標準正規分布からもとの分布を作ることが可能である。

標準正規分布には「基準正規分布について、その中心ゼロから、ある距離以上にある値が出現する確率 (p59)」をまとめた数表が作成されており、多くの統計学の教科書に付録として載っている。よって、

1. ある基準値を考える。
2. その基準値を正規化する。
3. 正規分布表を引く。
4. 「ある基準値」以下 (もしくは以上) の値が出現する確率がわかる！

しかしながら、現在ではコンピュータを使って正規分布曲線を簡単な操作によって計算できる。そのため、正規分布表を引くことはおろか、基準化という作業すら不要な場合がほとんどである。

m と σ の推定値

これまでは「母平均」や「母分散」を中心として議論を進めてきた。しかし、母集団のデータ全てが得られるようなケースはまずない。我々が得ることのできるのはたいていの場合は限られたデータからなる標本である。そこで、標本から得られる情報を用いて母集団の平均、分散を推定する方法が必要となる。

まず母平均の推定値であるが、これについては標本のデータを用いて計算した平均値、つまり標本平均が最もよい推定値であることが明らかとなっている。そのため、単純に標本平均を計算してこれを母平均の推定値であるとしてよい。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

ここで $\bar{x}$ が標本平均で、n が標本の個数、x_i が個々の標本のデータである。

次に母分散の推定値であるが、平均と同じように標本から分散を計算してそれを推定値とすればいいかといえば、そうではない。母分散の定義は次のようなものだった。

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \tag{6}$$

しかし、この式を標本に適応して計算される標本分散は母分散に対して若干小さくなる傾向がある[*2]。

[*1] ここで $\sum$ という記号は、記号以降の数式中にある記号の下で指定された添字 (ここでは i) を下で指定された数値 (ここでは 1) から記号の上で指定された数値 (ここでは n) まで 1 ずつ増やして全てのパターンの数式を作り、その総和を計算するという操作を略記したものと約束する。例えば $\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$ といった具合。

[*2] 標本平均の代わりに母平均を使えるのであれば偏りは生じない。ただしそのような状況はほとんどない。

この偏りを修正するには、割る数を少し小さくしてやればいい。

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 \tag{7}$$

この式により計算される値 s^2 を、母分散に対して偏りのない分散という意味で不偏分散と呼ぶ。また、その平方根 s を標準偏差と呼ぶ。

自由度について

母分散を計算するときに使った n、不偏分散を計算するときに使った $n-1$ は自由度と呼ばれる。これは単なるデータ数とは少し意味合いが違って、「ある系の位置を一意的に決定するのに必要な、独立な量の個数」と定義される[*3]。

「位置」というのは個々のデータ全ての値のことで、例えば n 個のデータの値を決定するには、当然 n 個のデータが必要となる。しかし、そこに「n 個のデータの平均値」という情報が与えられると、n 個のデータの値を n-1 個のデータから決定することができるようになる。n-1 個のデータがわかれば、残りのひとつは平均値と n-1 個のデータから計算できるためである。

この自由度という概念を用いて不偏分散の式を言葉で書き直せば次のようになる。

$$\text{不偏分散} = \frac{\text{偏差平方和}}{\text{自由度}} \tag{8}$$

偏差平方和というのは、平均値からの距離 (偏差) の平方 (二乗) の和という意味で、式 (7) からわかるように、次の式により計算される値である。

$$\sum_{i=1} n(x_i - m)^2 \tag{9}$$

[*3] ランダウ=リフシッツ「力学・場の理論」