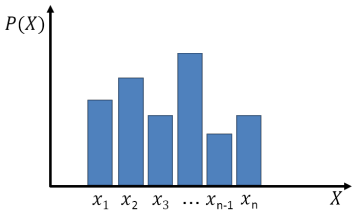
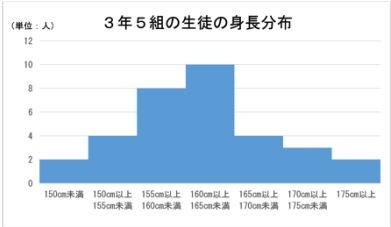
データ分布

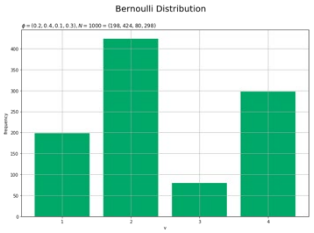
■度数分布図(ヒストグラム)



離散型の量的変数



連続型の量的変数

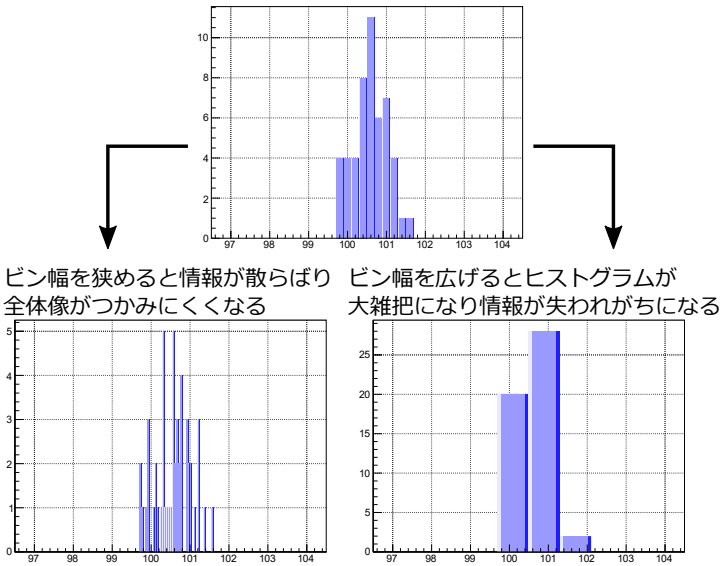


カテゴリ変数

ヒストグラムは、可視化でしかない。その読み取りは受け手側の判断に委ねられる。

■ビン幅

ビン幅が変わると情報の見え方が変わる。
統計ソフトでは、自動的にビン幅を決めて描画してくれる。

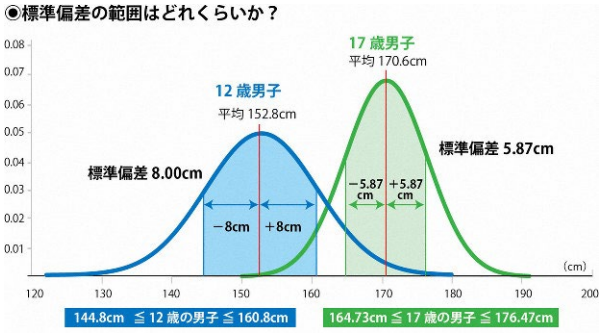


■統計量

　得られたデータに対して何らかの計算を実行して得られた値を統計量という。
データそのものの性質を記述し、要約するための統計量を、「記述統計量」または「要約統計量」という。
　記述統計量は、データの持つ情報のうち、捨てている情報がある。

身長
170
169
176
175
:

生データ



「平均値」、「標準偏差」
・・・記述統計量

◆代表的な記述統計量

・代表値：平均値、中央値、最頻値
・ばらつきを示す値：分散、標準偏差

◆平均値（mean）

　標本からの平均値「標本平均値」

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

◆中央値（median）

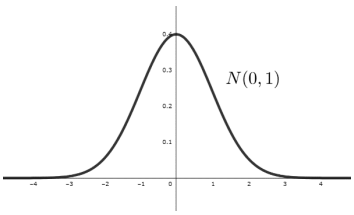
　大きさ順に並べたときの中央の値
要素が奇数個の場合は中央値はひとつ。偶数個の場合は中央値はふたつ。また、極端にばらつきがあっても影響を受けることが少ない。

◆最頻値（mode）

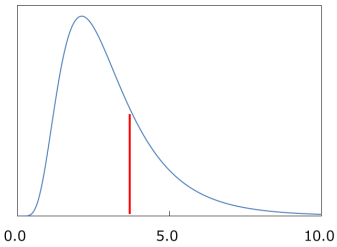
　もっとも頻繁に現われる数。
全体としての値の典型的な現われる把握できる。

■代表値のイメージ

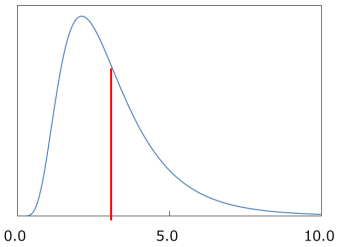
左右対称な山型だと、平均値、中央値、最頻値はだいたい一致する



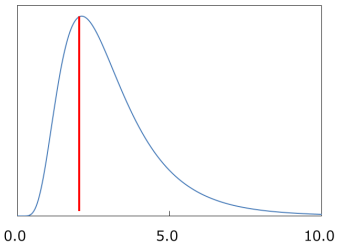
非対称の場合は、一致しない



平均値



中央値

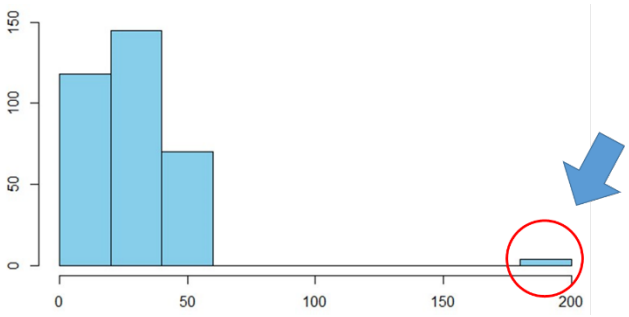


最頻値

◆外れ値（outlier）

極端に大きい値、小さい値を外れ値という。

外れ値は中央値には影響が少ないが、平均値には大きな影響を与える。



■分散と標準偏差

分散の幅、または、データのばらつきをとらえる。

「**分散**」（variance）または「**標準偏差**」（standard deviation、SD）

標本から統計量を求め、評価している場合を「標本分散」（sample variance）や「標本標準偏差」という。

標本分散 S_n^2 は、

$$s_n^1 = \frac{1}{n}\left\{(x_1 - \bar{x})2 + \cdots\right\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})2$$

「標本分布」の性質は、

- S_n^2 ＞＝0
- 全ての値が同一である時 0
- ばらつきが大きいと、 S_n^2は大きくなる

標本標準偏差は、標本分布のルートを取った値

$$s_n^2 = \sqrt{S_n^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})2}$$

分布と標準偏差は√を取るかどうかの違いだけ。

標準偏差は、平方根を取っているので,元の単位と同じ。感覚的に分かりやすい。