

データの分布と 統計量の分布

60161017 金城 清史

資料

- ・ 医薬研究者の視点からみた道具としての統計学
奥田千恵子 著

- ・ 5章「データの分布」

- ・ 6章「統計量の分布」



5章 「データの分布」

- 2項分布
- 大数の法則
- 正規性の検定
- 少数データ

5章 「データの分布」

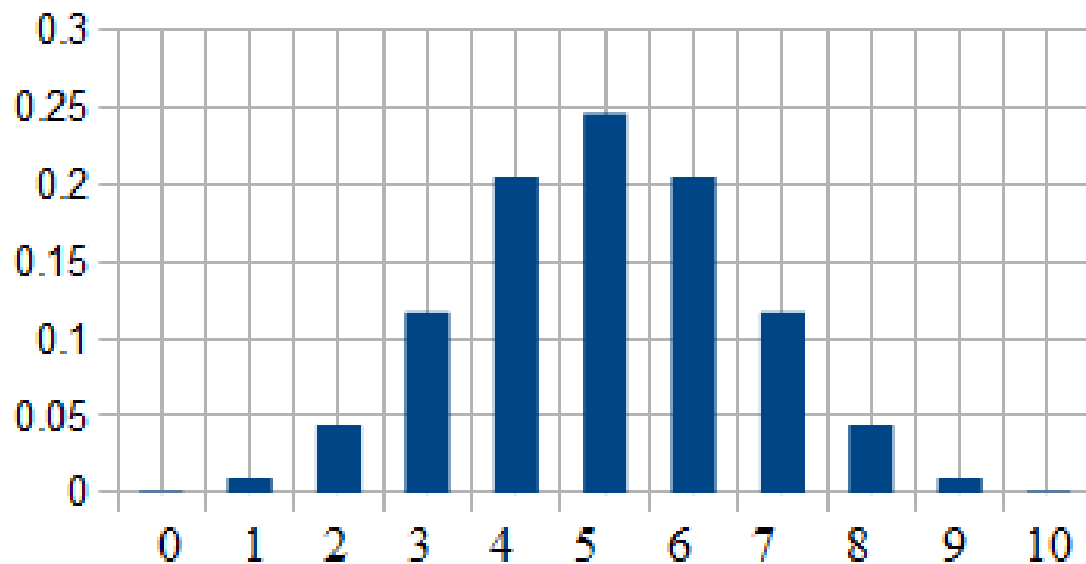
- 2項分布
- 大数の法則
- 正規性の検定
- 少数データ

2項分布

- コインの裏, 表のような可能性が2つしかない時の確率分布
- 表が出る確率を p , 裏がでる確率を $(1-p)$ とすると

$$P(x) = {}_{10}C_x \times p^x (1-p)^{10-x} = {}_{10}C_x \times \left(\frac{1}{2}\right)^{10}$$

x	${}_{10}C_x$	P(x)
0	1	0.000976563
1	10	0.009765625
2	45	0.043945313
3	120	0.1171875
4	210	0.205078125
5	252	0.24609375
6	210	0.205078125
7	120	0.1171875
8	45	0.043945313
9	10	0.009765625
10	1	0.000976563



大数の法則

- 歪みのないコインを投げ続ければ, その確率分布は先験的に規定された確率 $p=\frac{1}{2}$ の2項分布に近づく
 - ⇒ 十分な標本数の集団を調べれば, その集団内の傾向は, その標本が属する母集団の傾向と同じになる.
- 大数の法則

母平均 μ , 母分散 σ^2 をもつ母集団から, 大きな n は標本 $X_1, \dots, X_n$ を抽出し, その標本平均を $\bar{X}$ とすると $\bar{X}$ は μ に収束する

$$\lim_{n \rightarrow \infty} P\{|\bar{X} - \mu| \geq \varepsilon\} = 0 \text{ が成り立つ}$$

正規性の検定法

- 正規性の検定
 - データが正規分布に従わないか,従わないとはいいきれないかの検定
- 正規性の検定法
 - 正規格率紙
 - コロモゴロフ・スミルノフの適合度試験
 - リリーフォースの検定
 - シャピロ・ウィルクスの検定

正規分布前提の統計解析法

- 母集団が正規分布に従うことを前提にした統計解析法
 - t 検定(9章)
 - 直線回帰など

⇒ 正規分布に従うことを前提にしている検定なので
注意

正規性検定

- パラメトリック検定
 - 対象がある分布に従うとき
 - 制限(前提条件)がある
- ノンパラメトリック検定
 - サンプル数が少ないとき
 - 検定力が低下する

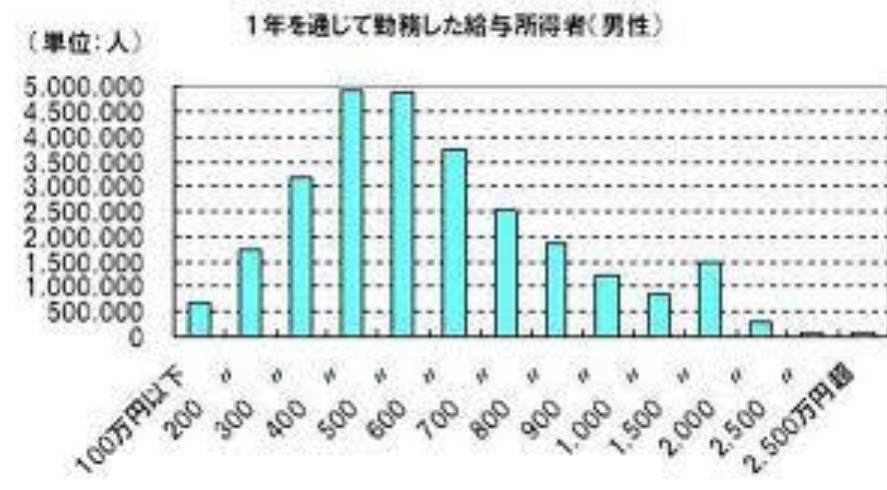
どちらで行うかを決めることが重要

ノンパラメトリックな検定法(12章)

- データの分布の形を問題にしない検定法
 - パラメトリック手法と比べて、母集団の分布などの前提を必要としない
- マン・ホイットニーの検定など

少数データではグレーゾーン？

- 通常の実験研究のデータ数程度では,上のいずれの方法を用いても正規性の検定は事実上不可能？
- ほとんどの場合,正規性が否定されることはない
- 実用的にはデータをヒストグラムなどにして視覚的に確かめるしかない.
- 明らかな分布の歪みや外れ値がない限り,正規分布を前提としている統計解析方法を用いる



6章 「統計量の分布」

- 記述統計, 推測統計
- 中心極限定理
- 標準誤差
- 標準正規分布
- t分布

6章 「統計量の分布」

- 記述統計, 推測統計
- 中心極限定理
- 標準誤差
- 標準正規分布
- t分布

記述統計・推測統計

- 記述統計(descriptive statistics)
 - 研究対象の集合としての性質を様々な統計量(statistic)を用いて,整理,記述する.
 - 観察された集団のみを扱う
 - 全数調査,大標本で行われる.
- 推測統計(inferential statistics)
 - 1回限りの調査や少数の実験データの統計量から,不確実さの度合を数量化した上で,母集団の特性を推定する.
 - 観察された集団を母集団から抽出された標本とみなし,母集団の諸性質を推論する方法

中心極限定理

標本数が大きければ,母集団の分布型にかかわらず
標本平均値の分布は正規分布に近似的に従う

- この定理の条件はかなりゆるやか
 - 標本数が10～20程度でも成り立つ

点推定

- 母集団の分布の形が分かっている、その母数(例: 母平均値)がわからない時は、標本から得られた1つの値(例: 標本平均)によってその母数を推定することを **点推定** という。
- 例) 5人のテストの点数の平均値が60点のとき、母集団のテストの点数の母平均値を推定せよ
 - A. 60点

標準誤差

- 一般的な定義
- 任意の統計量の標本分布の標準偏差
 - 中央値,標準偏差,比率など,どのような統計量でもそれぞれの標本分布について標準偏差を求めることができる

平均値の標準誤差 $= \frac{\sigma}{\sqrt{n}}$ で表される.

	A	B	C
1	55	65	60
2	65	70	75
3	67	65	45
4	53	60	42
5	60	47	78
6		55	64
7		55	56
8		63	90
9		55	30
10		65	60
平均値	60	60	60
標準偏差	6.082763	5.953524	11.04536

例)

A:標準偏差=6.1,標本数=5のとき, $\frac{6.1}{\sqrt{5}} = 2.73$

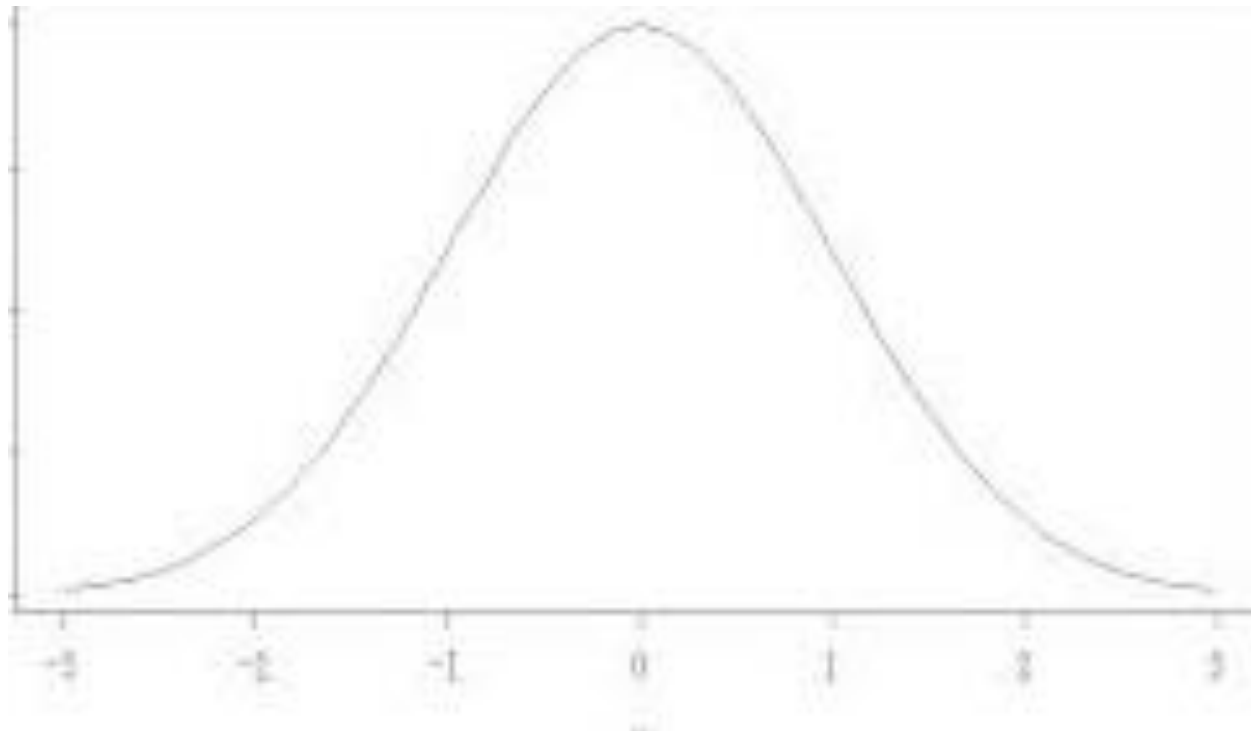
B:標準偏差=6.0,標本数=10のとき, $\frac{6.0}{\sqrt{10}} = 2.00$

C:標準偏差=11.0,標本数=10のとき, $\frac{11.0}{\sqrt{10}} = 3.48$

標準正規分布

正規分布, $N(\mu, \sigma^2)$ に従う母集団に属するデータ, x を標準化した z 値は, 平均値が 0, 標準偏差が 1 の正規分布に従う. これを標準正規分布と呼ぶ.

略して $N(0, 1)$ $\Rightarrow$ 分布表を利用できるなどの利点



t分布

- 少標本の場合,標本正規分布に当てはまりにくい
- t分布
 - 標本数が100近くになると,t分布もほぼ標準正規分布とほぼ重なる.

まとめ

- パラメトリック・ノンパラメトリックな検定のどちらを使うべきかを判断する.
- 推定された値がどのくらい信頼できるものかを判断できるようにする
- 分布を利用できるようにする(有意差検定(8章))

終わり

4章の復習？

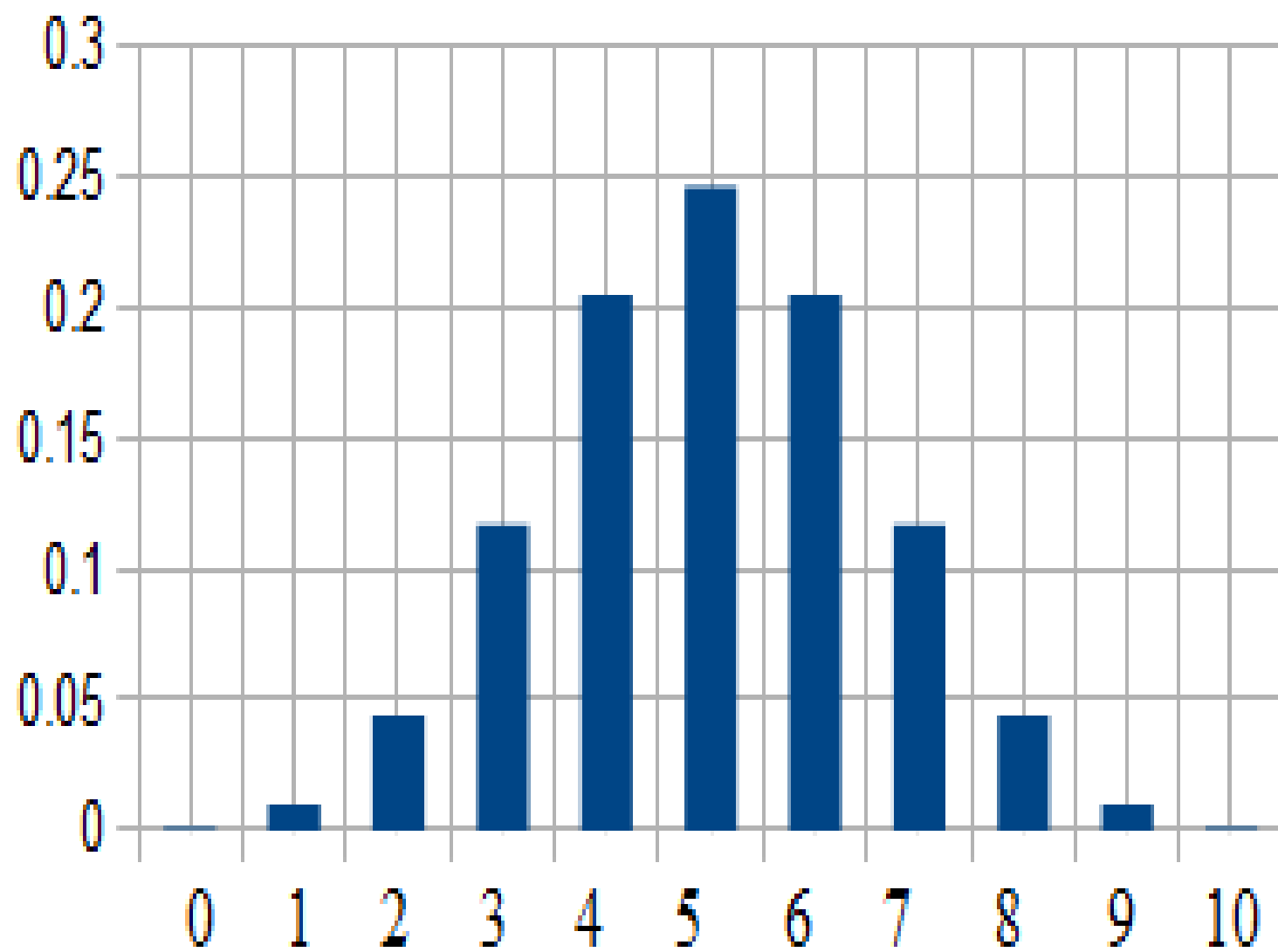
- 正規分布

- 偶然誤差の性質

1. 正負同程度に起こる
2. 絶対値の大きさは誤差が大きいほど生じにくい
3. データの平均値が真の値の推定値となる



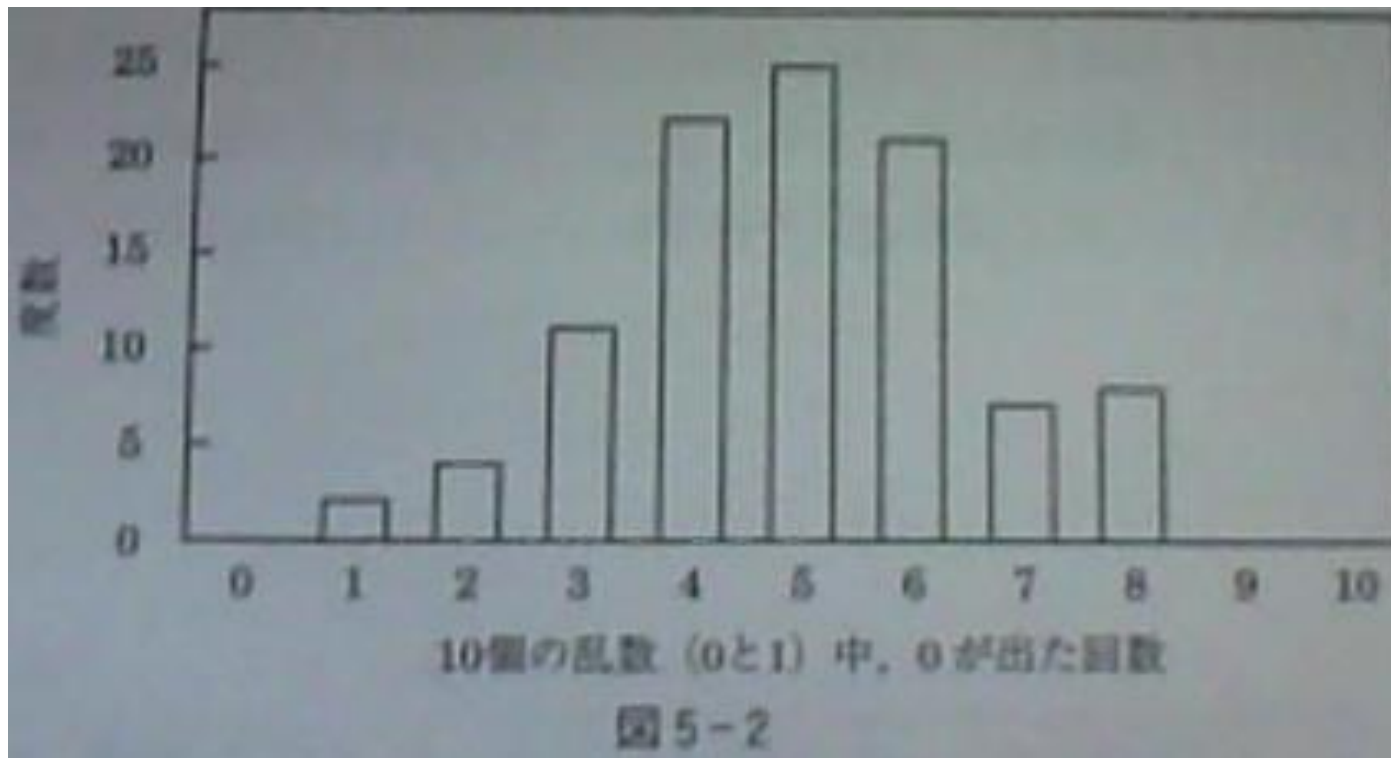
この3つの条件を満たす論理的確率分布を**正規分布**という



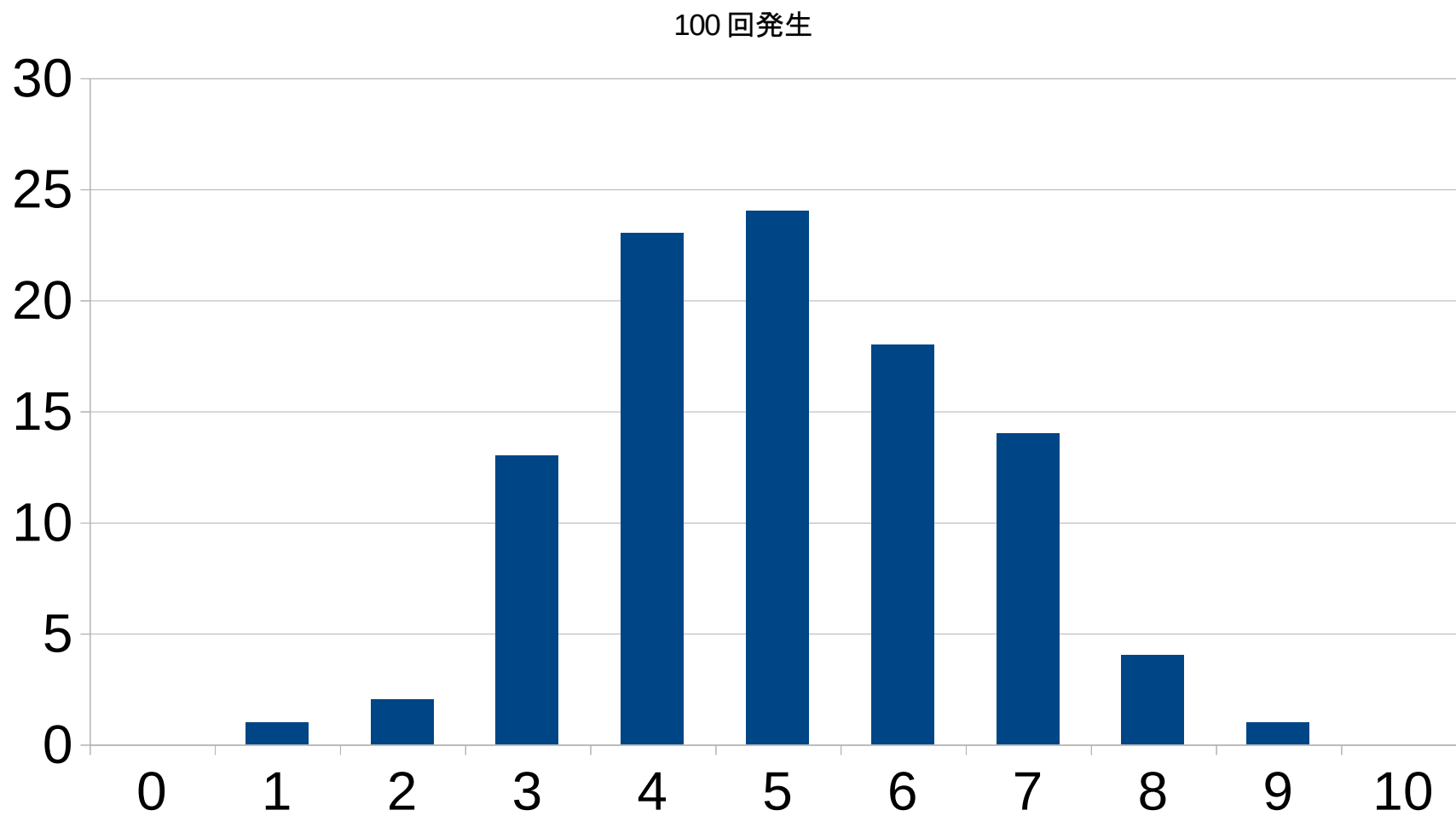
コンピュータを用いた実験

- 10個のコインを投げて行う代わりに乱数を用いて0か1を10個ずつ発生させて0が出た回数をグラフに表す.

(図5-2 100回発生させた結果)

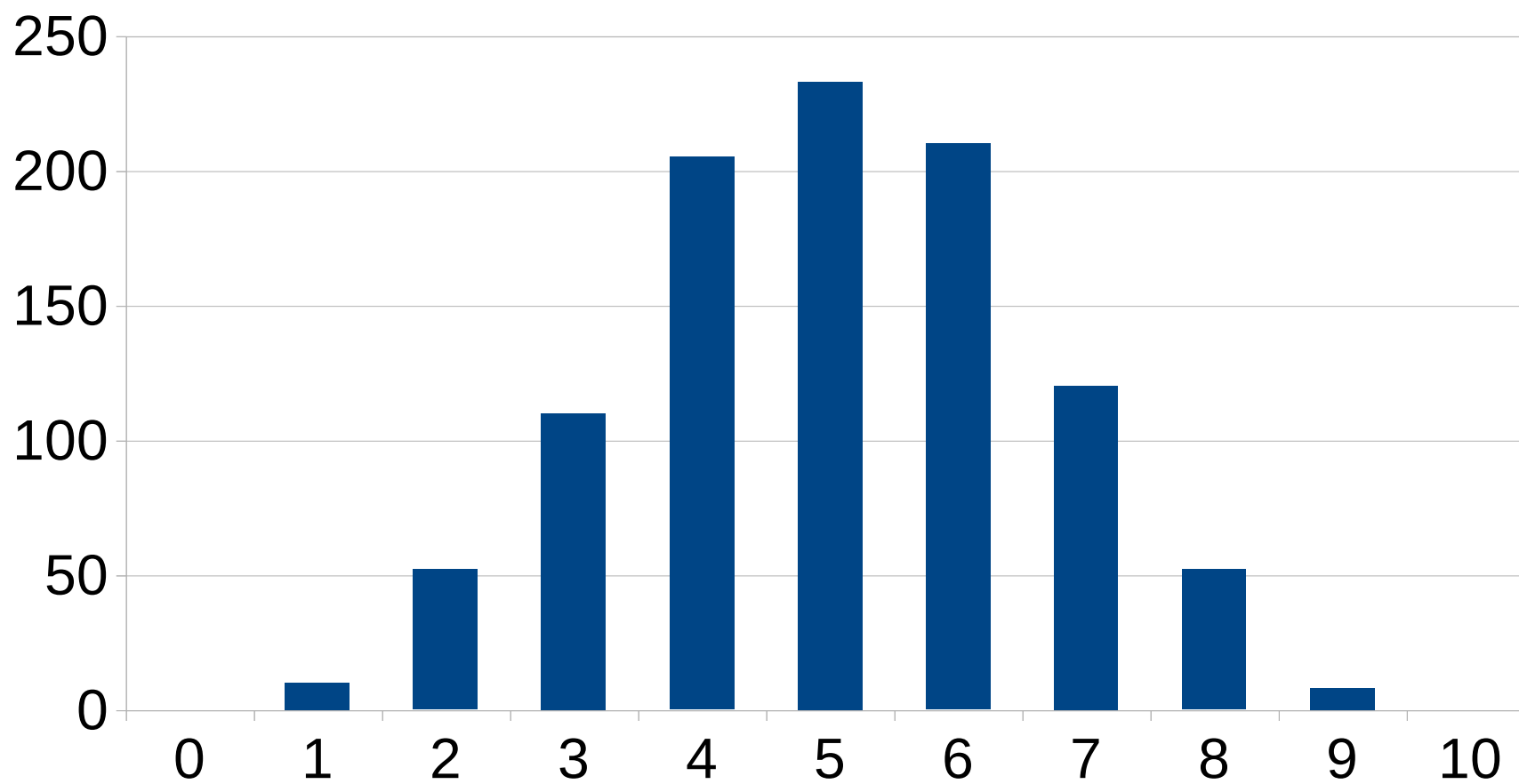


100回発生



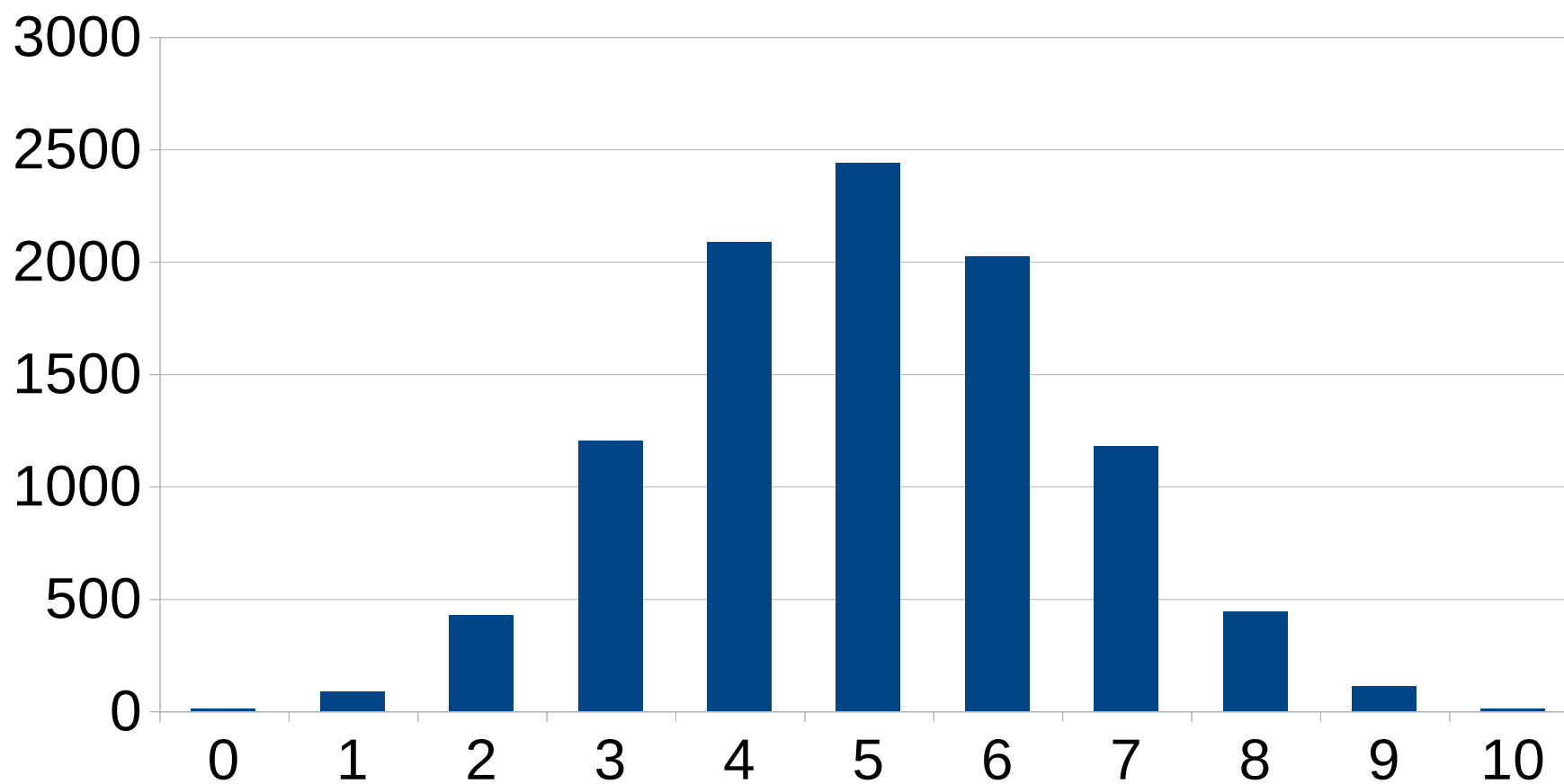
1000回発生

1000 回発生



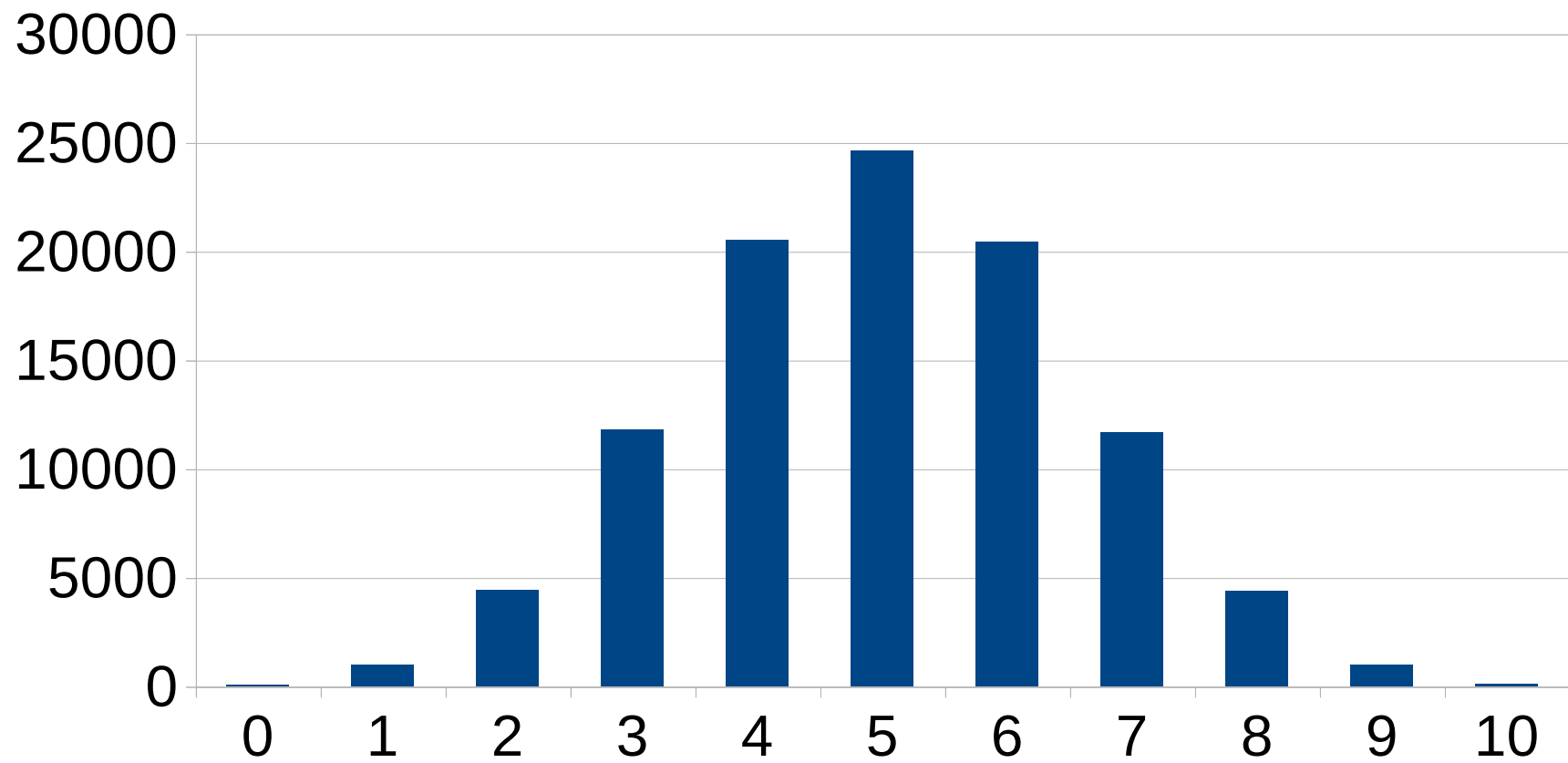
10000回発生

10000 回発生



100000回発生

100000 回発生



正規分布のパラメータ

- データが近似的に正規分布に従うことが確認されれば、母集団が正規分布していると仮定できる

母集団		標本	
母平均値	μ	標本平均値	$\bar{x}$
母標準偏差値	σ	標本標準偏差値	SD

正規分布は以下の式を用いて、釣り鐘型の中心位置を表す平均値とばらつきの程度を表す標準偏差の2つで表すことができる。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

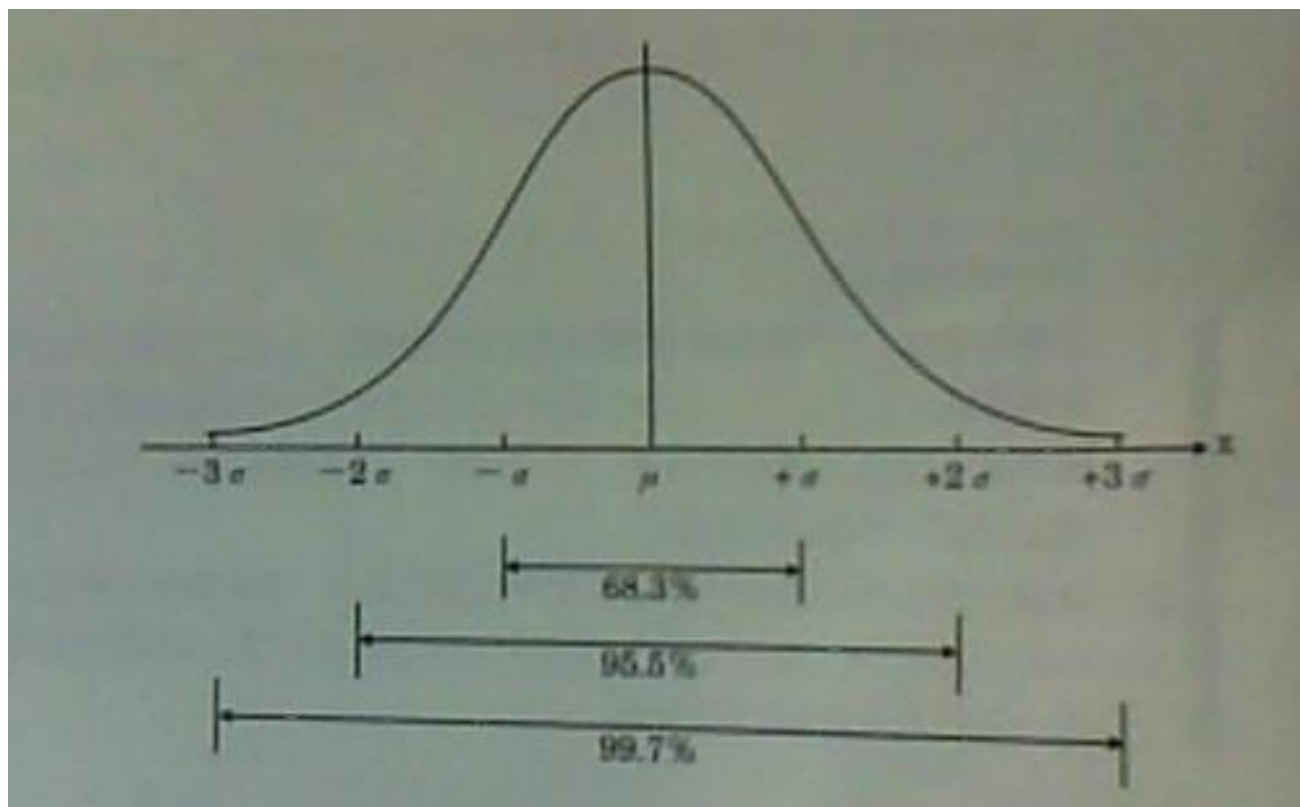
略して $N(\mu, \sigma^2)$ と表せる

$$(-\infty < x < \infty)$$

平均値と標準偏差から全体を知る

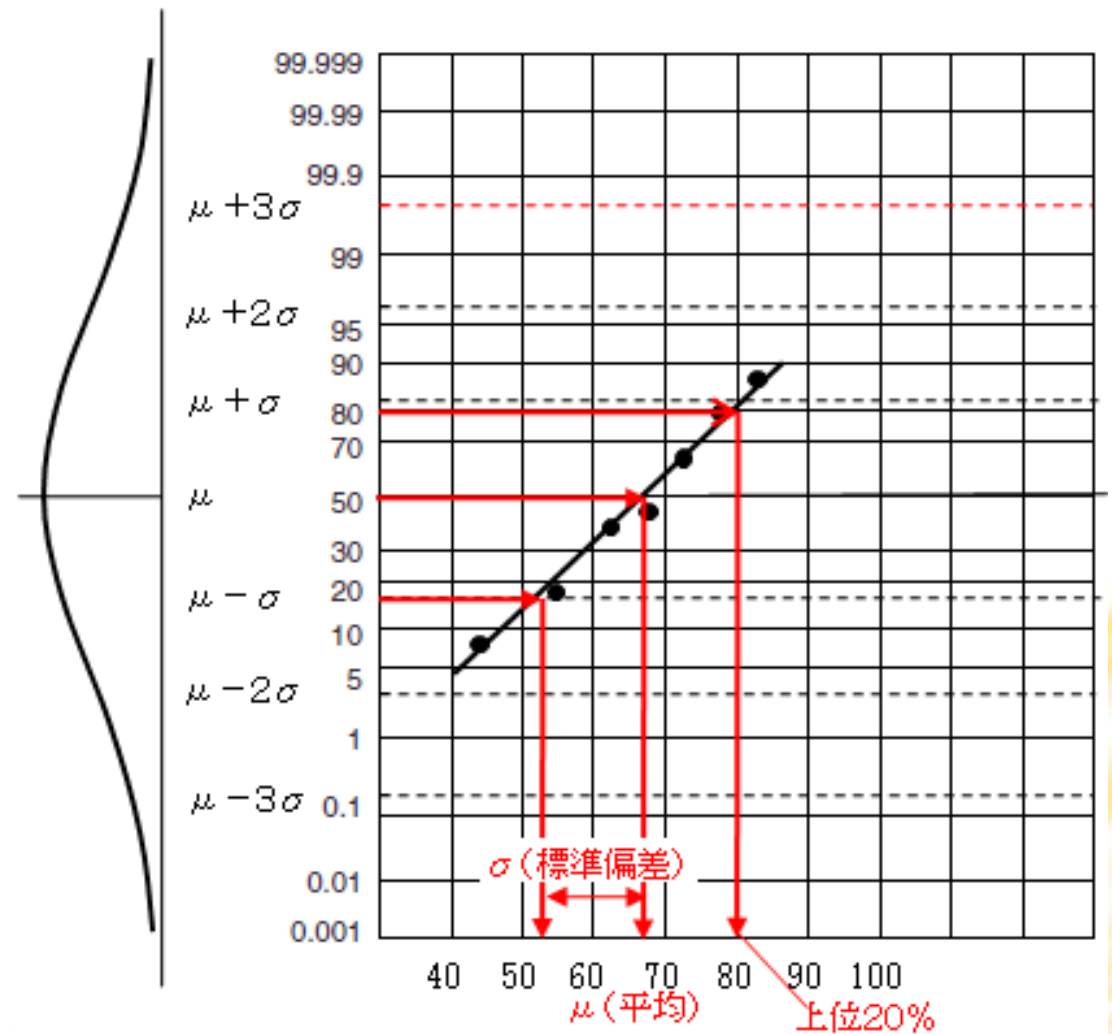
正規分布曲線で

「対象の年齢は 45 ± 8 歳」である場合分布は
 45 ± 8 歳の範囲には全体の68.3%,
 $45 \pm 8 \times 2$ 歳の範囲には95.5%が含まれてる



正規確率紙

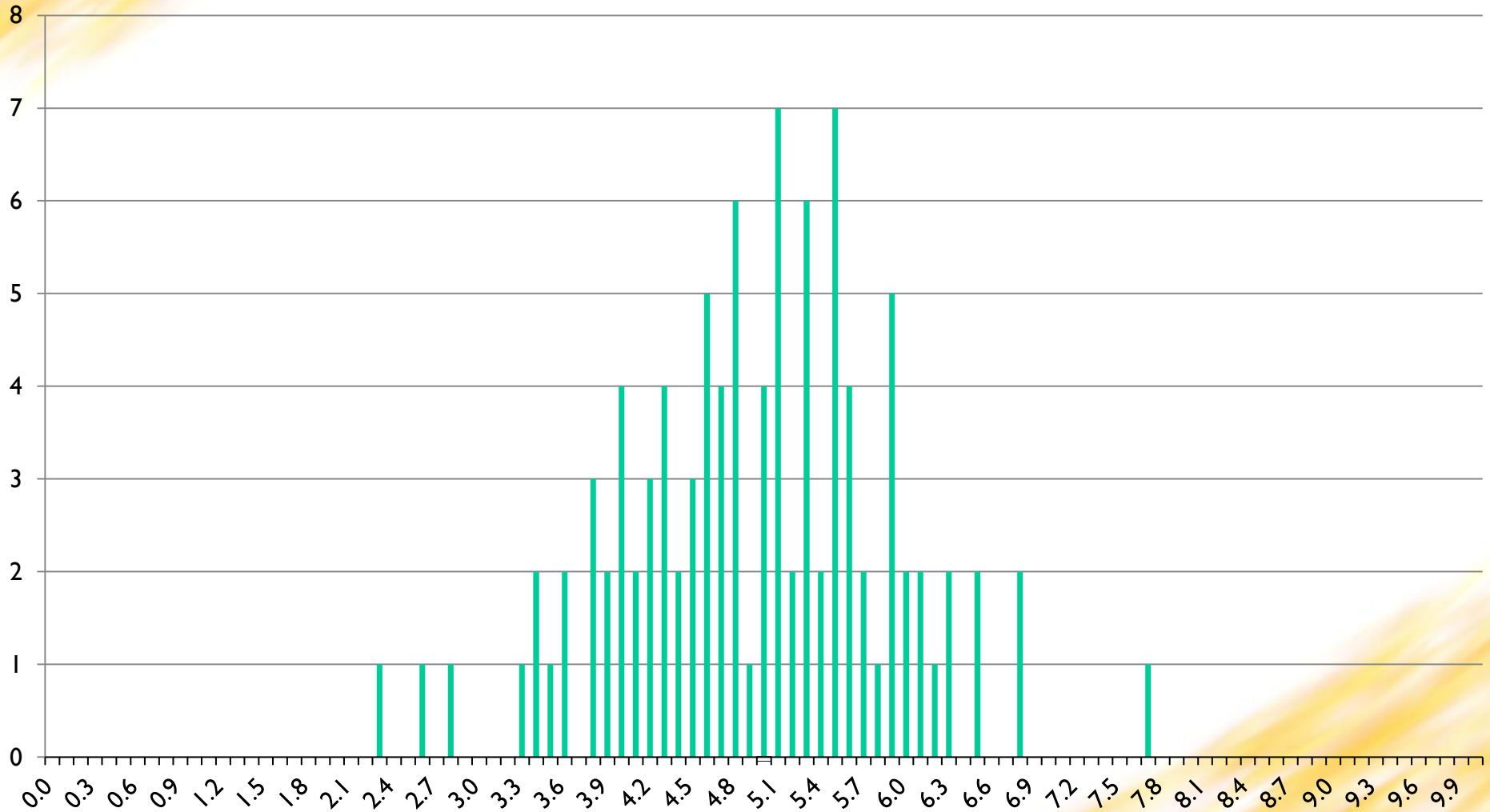
- 点が直線に近い⇒正規分布
- 標準偏差
 - $\sigma = 66 - 53 = 13$



コルモゴロフ・スミルノフの 適合度試験

- **コルモゴロフ・スミルノフの適合度試験**
- 仮説検定の1種
 - 有限個の標本に基づいて、二つの母集団の確率分布が異なるものであるかどうか、あるいは母集団の確率分布が帰無仮説で提示された分布と異なっているかどうかを調べるために用いられる
- **リリーフォース検定**
 - コルモゴロフ・スミルノフの適合度試験の改良版
- **シャピロ・ウィルクスの検定**
 - 仮説検定の1種

中心極限定理(一樣分布) 100回



標準正規分布

- 標準化

- 異なる基準で測定された値が比較できると便利だが、異なるデータを一定の基準に揃える

母平均値, μ ,および,母標準偏差, σ の正規分布に従う母集団に属する変数 x を,平均値を原点,0に移し,標準偏差が1になるように変換したものを z 値という.

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

標準正規分布の利用

40人のクラスの平均点が60点で、ある5人の平均点が66点のとき、ある5人は学校の平均点と比べて差があるか？(標準偏差を18.0とする)

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{66 - 60}{\frac{18}{\sqrt{40}}} \doteq 2.108$$

※通常の有意差検定では5%
以下の場合、有意差ありとする
⇒ **差がある**と考える

