

統計的推定の基礎 (母集団と標本)

キーワード：

記述統計学・推測統計学・母平均・母分散・標本平均・標本分散・不偏分散・偏差・偏差値・確率密度関数・(標準)正規分布・中心極限定理・標準化・z 値

標本抽出¹

母集団	→	標本
(母数 ² =パラメータ)		(統計量 ³)
本当に知りたいもの		標本から計算できるもの
母平均		標本平均
母分散		標本分散／不偏分散
母標準偏差		標本標準偏差

母集団のデータを分析する調査方法：

「全数調査 (complete survey)」・・・母集団のすべてのデータを調査する
「標本調査 (sample survey)」・・・母集団の一部のデータを調査する

無作為抽出によって、母集団の特徴を確率で客観的に推測することができる

記述統計学⁴・・・観察対象となるデータの特徴をひとつの数値にまとめること
(平均値・中央値・分散・標準偏差・最大値・最小値など)

推測統計学⁵・・・「推定」(点推定と区間推定)と「仮説検定」から構成
母集団から抽出したサンプル(標本)に基づいてその母集団
全体の特徴や性質を推測(infer)しようとする

¹ 標本抽出のことを一般的に無作為抽出 (random sampling)という

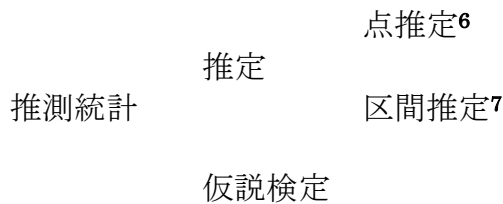
² 母数 (population parameter)

³ 統計量 (statistic)

⁴ 記述統計学 (descriptive statistics)

⁵ 推測統計学 (inference statistics)

推測統計の分類



点推定：

- ・ サンプルから得た「統計量」を用いて母集団の「パラメータ」を推定する
- 例) 「日本の男子大学生の平均身長は 170cm くらいだろう」

区間推定：

- ・ 点推定で推定した「パラメータ」のばらつきや「信頼区間」⁸を示す
- 例) 「日本の男子大学生の平均身長は 160～170cm くらいだろう」

仮説検定：

- ・ 区間推定値を使って、母集団が特定の分布に従っているかどうか検証する。
- ・ サンプルから得た「統計量」が特定の分布に従う母集団から抽出されたという仮説（＝帰無仮説）をたて、その仮説を検証する。

帰無仮説から予想される「統計量」とサンプルから抽出された「統計量」が一致する確率（＝ p 値）を計算し、その確率が有意水準（10%, 5%, or 1%）よりも小さい場合には「統計的な有意差がある」として帰無仮説は棄却される。

仮説検定方法

パラメトリック⁹な検定手法・・・データが特定の確率分布に従うことを仮定

- ・ F 検定
- ・ t 検定

ノンパラメトリック¹⁰な検定手法・・・特定の確率分布を仮定しない

- ・ カイ二乗検定
- ・ Wilcoxon 検定
- ・ フィッシャーの正確確率検定

⁶ 点推定 (point estimation)

⁷ 区間推定 (interval estimation)

⁸ 具体的には「95%信頼区間」「99%信頼区間」などが用いられる

⁹ 母集団の分布を特定し、測定値として連続量を想定する検定。母集団の分布として正規分布が想定されることが多い（武藤著『統計解析ハンドブック』1995年）。

¹⁰ 母集団の分布を特定せずに、測定値として非連続量を想定しない検定。

1. 母集団と標本における平均と分散（パワポ講義）

次のデータは高田馬場駅前発、早稲田大学正門前行きバスの実際の発車時刻を 5 日間記録したものである。

	12/1	12/2	12/3	12/4	12/5	平均発車時刻
09:00 発	08:59	08:59	09:00	09:01	09:01	09:00
12:00 発	11:56	11:57	12:01	12:02	12:04	12:00

平均発車時刻 09:00 からのデータの離れ具合を調べると

	12/1	12/2	12/3	12/4	12/5	平均
09:00 発	-1	-1	0	1	1	0
12:00 発	-4	-3	1	2	4	0

推測統計として分析する場合（＝母集団を推計する場合）

$$u_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

記述統計として分析する場合（標本＝母集団の場合）

母集団の分散 σ^2 （ σ 二乗）は次の式で求めることができる。

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}.$$

	nine	noon	
	-1	-4	
	-1	-3	
	0	1	
	1	2	
	1	4	
average	0	4	
stata	1	3.39	標本標準偏差(不偏分散の平方根)u
EXCEL(STDEV)	1	3.39	標本標準偏差(不偏分散の平方根)u
EXCEL(STDEVP)	0.75	3.03	母標準偏差 σ

RStudio 上で計算してみる

T.Baba - Waseda Bus

```
nine <- c(-1, -1, 0, 1, 1)
noon <- c(-4, -3, 1, 2, 4)
```

分散を求める

```
var(nine)
```

```
## [1] 1
```

```
var(noon)
```

```
## [1] 11.5
```

標準偏差を求める

```
sd(nine)
```

```
## [1] 1
```

```
sd(noon)
```

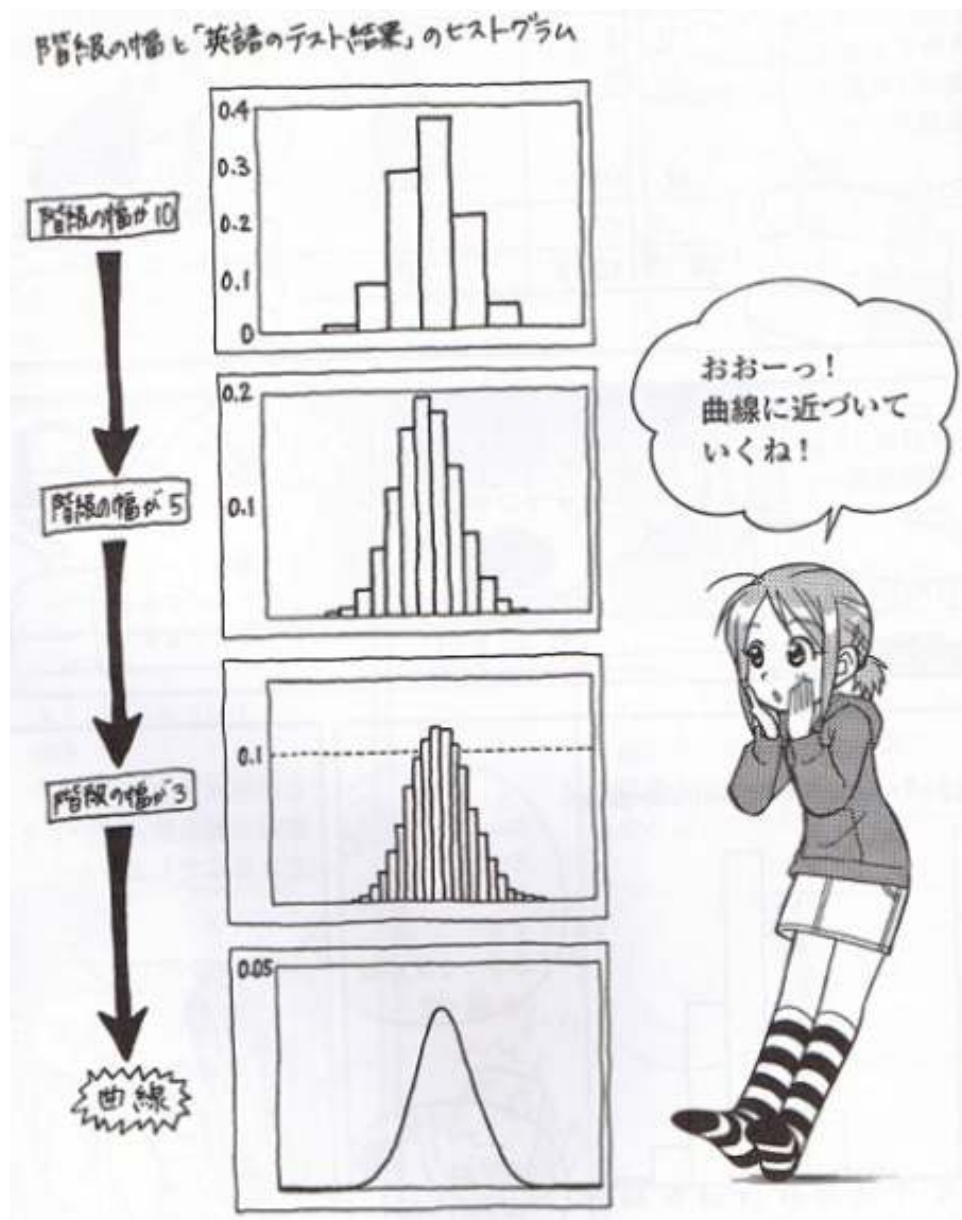
```
## [1] 3.391165
```

詳細は次のサイトを参照

http://www.ner.takushoku-u.ac.jp/masano/class_material/waseda/keiryo/4_descriptive_stat.html#母分散と不偏分散

2. 確率密度関数・正規分布・標準正規分布

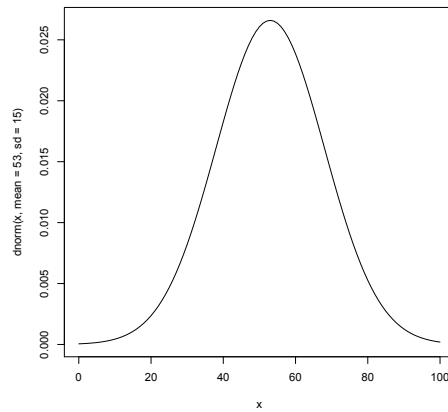
確率密度関数・・・ヒストグラムにおける階級の幅を極限まで狭めた曲線の式



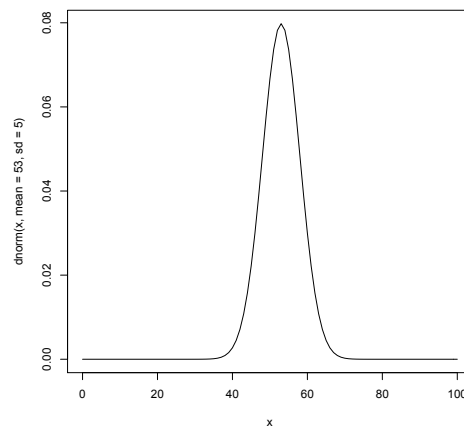
正規分布 (Normal Distribution = bell curve)
ド・モアブル¹¹が二項分布の近似として発見した確率分布

¹¹ アブラーム・ド・モアブル (Abraham de Moivre, 1667- 1754) はフランスの数学者。

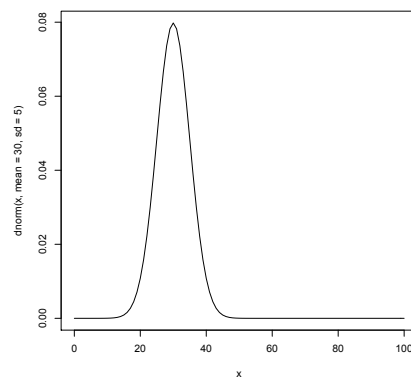
平均が 53 で標準偏差が 15 の正規分布



平均が 53 で標準偏差が 5 の正規分布



平均が 30 で標準偏差が 5 の正規分布



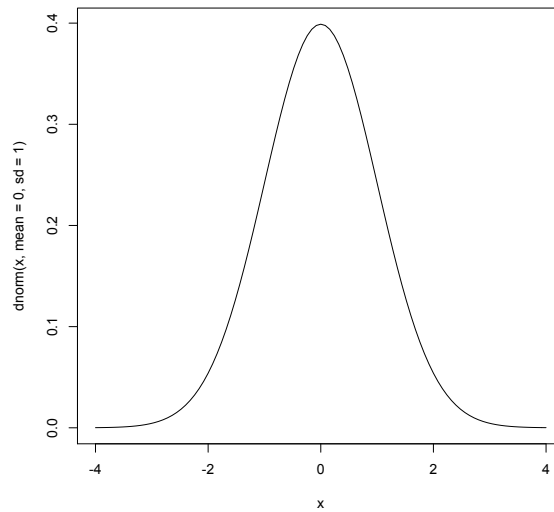
特徴：

- ①平均を中心に左右対称である
- ②平均と標準偏差の影響を受ける

Useful Commands on R

How to draw a Normal Distribution

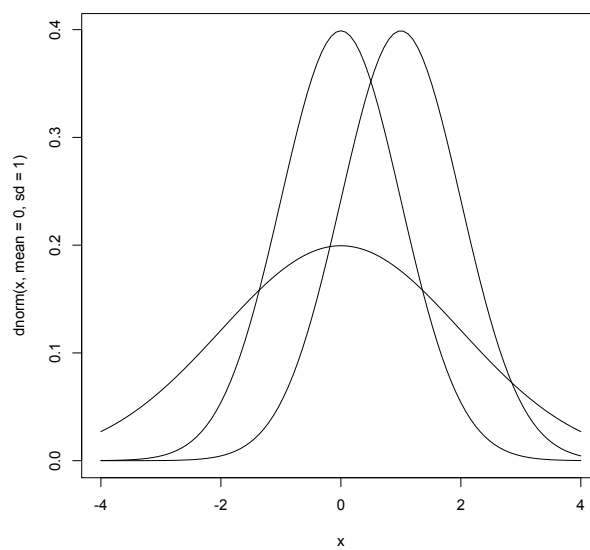
> curve(dnorm(x, mean=0, sd=1), from=-4, to=4) • • • 標準正規分布



When you want to add lines to the existing graph.

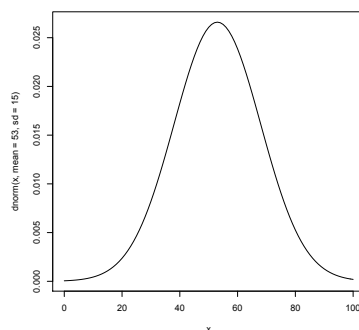
> curve(dnorm(x, mean=0, sd=2), add=TRUE)

> curve(dnorm(x, mean=1, sd=1), add=TRUE)

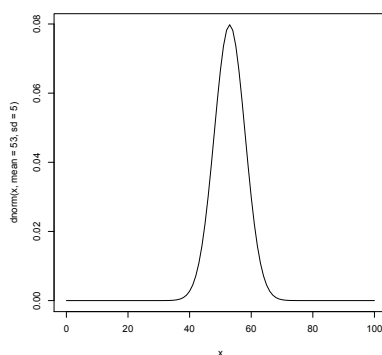


練習問題：次のグラフをRを使って描いてみよう
(コマンドは脚注にあります)

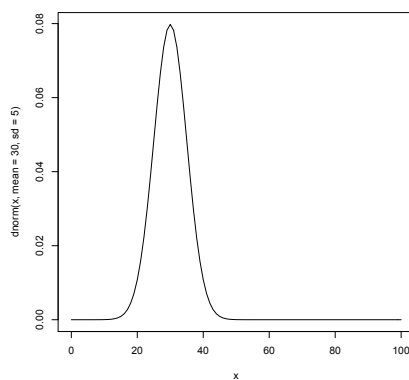
平均が 53 で標準偏差が 15 の正規分布¹²



平均が 53 で標準偏差が 5 の正規分布¹³



平均が 30 で標準偏差が 5 の正規分布¹⁴



¹² `curve(dnorm(x, mean = 53, sd = 15), from = 0, to = 100)`

¹³ `curve(dnorm(x, mean = 53, sd = 5), from = 0, to = 100)`

¹⁴ `curve(dnorm(x, mean = 30, sd = 5), from = 0, to = 100)`

正規分布が統計学上特別な地位を持つのは中心極限定理が存在するため

中心極限定理

X が平均 μ 、標準偏差 σ のある分布に従うならば、大きさ n の無作為標本に基づく標本平均 \bar{X} は、 n が無限に大きくなる時、平均 μ 、標準偏差 σ / \sqrt{n} の正規分布に近づく。

母分散・不偏分散・標本分散の関係

標本分散 (s^2) は母分散 (σ^2) の推定値にはならないことは統計学的に証明されている。母分散の正しい推定値は、標本分散 (s^2) ではなく不偏分散 (u^2)。

上記に関する詳細は、下サイトの「標本分布」の項目を参照

http://www.ner.takushoku-u.ac.jp/masano/class_material/waseda/keiryo/5_infer_stat.html#中心極限定理

正規分布 (Normal Distribution)のつづき

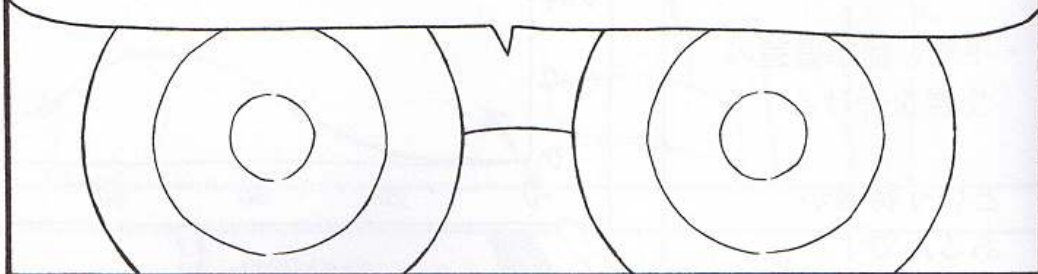
x の確率密度関数が

先ほどの式

$$f(x) = \frac{1}{\sqrt{2\pi} \times x \text{ の標準偏差}} e^{-\frac{1}{2} \left(\frac{x - x \text{ の平均}}{x \text{ の標準偏差}} \right)^2}$$

であるならば

『 x は、平均が^{うんたら}〇〇で標準偏差が^{かんたら}××の正規分布にしたがう』と、統計学では表現します！



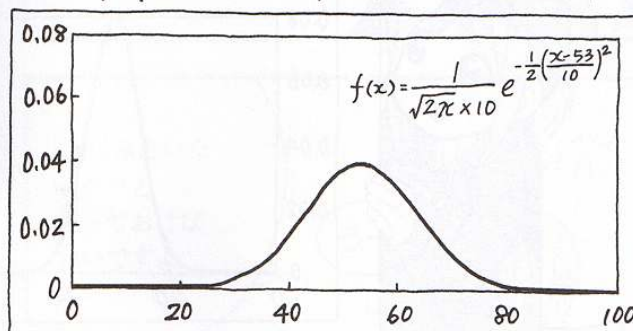
$e = 2.7182$ (自然対数の底)

先ほどのテストの例で
いきますよ

もし
「英語のテスト結果」の
確率密度関数が
右のものだった
ならば…



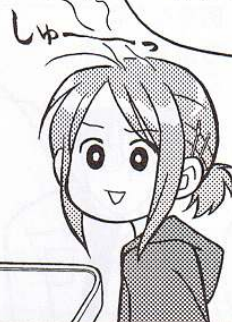
平均が53で標準偏差が10の正規分布



『「英語のテスト結果」は、
平均が53で標準偏差が10の
正規分布にしたがう』と
表現するわけです



な、
なるほどーッ



標準正規分布 (Standard Normal Distribution)

正規分布に変数変換を施した (=標準化した) 後の分布

変数変換: $z = (\text{個々のデータ} - \text{平均}) / \text{標準偏差}$

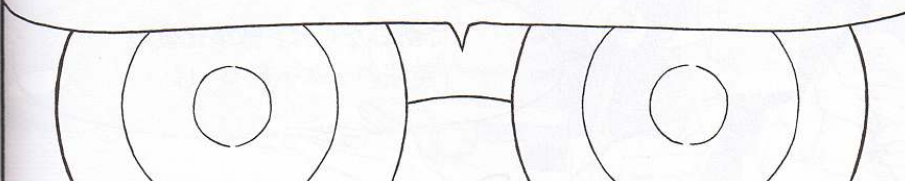
Z 値を求める

→ 標準正規分布表とよばれる変数に対応した確率をあらわす一覧表を用いて、コンピュータを使うことなく正規分布に従った事象の確率を求める事ができる

x の確率密度関数が

$$f(x) = \frac{1}{\sqrt{2\pi} \times x \text{ の標準偏差}} e^{-\frac{1}{2} \left(\frac{x - x \text{ の平均}}{x \text{ の標準偏差}} \right)^2} = \frac{1}{\sqrt{2\pi} \times 1} e^{-\frac{1}{2} \left(\frac{x-0}{1} \right)^2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$


であるならば
『 x は、平均が0で標準偏差が1の正規分布にしたがう』ではなく
『 x は、標準正規分布にしたがう』と、統計学では表現します!



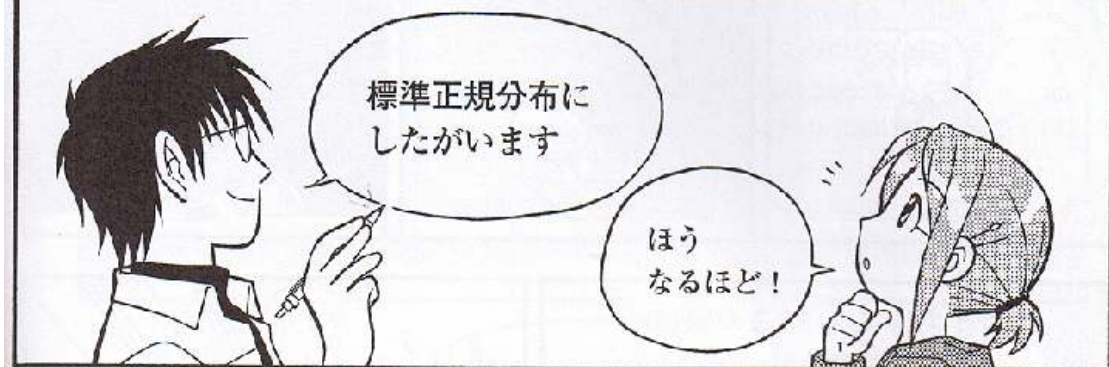
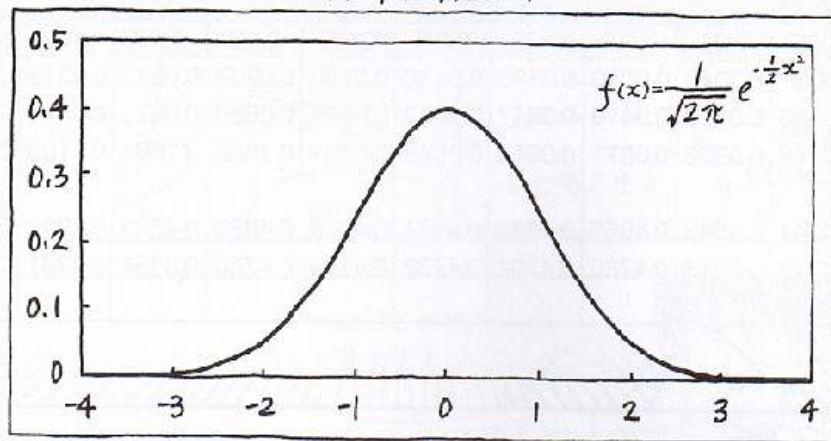
英語のテスト結果		英語のテスト結果 (標準化後)
生徒1	42	-1.1
生徒2	91	3.8
⋮	⋮	⋮
生徒10421	50	-0.3
平均	53	0
標準偏差	10	1

$$\frac{\text{個々のデータ} - \text{平均}}{\text{標準偏差}} = \frac{50 - 53}{10} = \frac{-3}{10} = -0.3$$

であるならば、標準化後の「英語のテスト結果」は…



標準正規分布



標準正規確率表 (Z に対する原点からの累積確率を求める表)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

練習問題 1 :

高校 1 年生全員がある予備校の数学のテストを受けました。採点したところ「数学のテスト結果」は平均が 45 点で標準偏差が 10 の正規分布に従うとみなせることがわかりました。

- (1) 70 点以上得点した受験生の割合は？・・・P (個々のデータ ≥ 70)
- (2) 50 点以上得点した受験生の割合は？・・・P (個々のデータ ≥ 50)
- (3) 40 点以下得点した受験生の割合は？・・・P (個々のデータ ≤ 40)
- (4) トップ 5%に入るためには何点必要か？

練習問題 2 :

缶コーヒー 200 個の内容量の平均値が、150.3g、標準偏差が 5g の正規分布に従うものとする。このとき次の間に答えよ。

(1) 内容量の多い方から 50 番目の缶の内容量はどれくらいか。

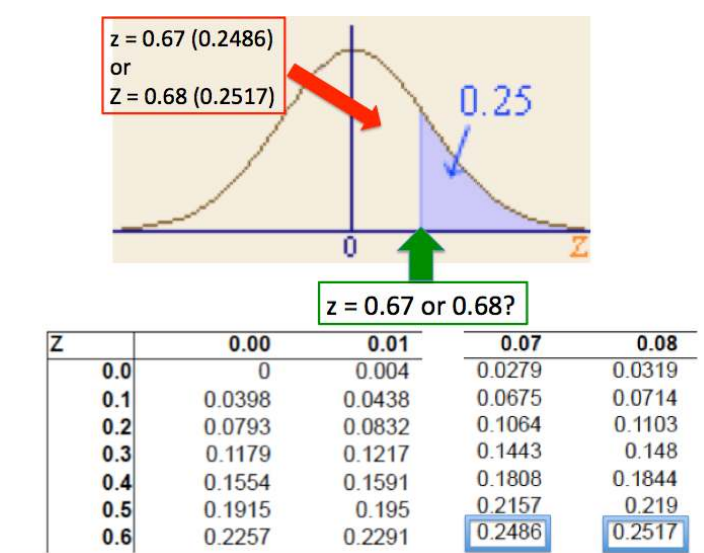
(2) 内容量の多い方から 5% の重さはどれくらいか。

[解答]

(1) $z = (\text{個々のデータ} - \text{平均}) / \text{標準偏差}$ なので、

$z = (\text{個々のデータ} - 150.3) / 5$ とおくと、 z は「平均値 0、標準偏差 1 の標準正規分布に従う」¹⁵。

「内容量の多い方から 50 番目の缶」とは 200 個の缶コーヒー中、内容量の多い順から数えて四分の一（面積が 0.25）という意味だから、下記の標準正規分布の右端の面積が 0.25 (50 / 200) を満たす z 値を求めればよい。



これを満たす z 値は、正規分布表より 0.67

その理由：

$z = 0.68$ ($z = 0$ から $z = 0.68$ までの面積 = 0.2517) だと右側の紫色の部分の面積 < 0.25

→ 「内容量の多い方から 50 番目 (= 0.25%)」を求めることができなくなる)

$z = (\text{個々のデータ} - 150.3) / 5$ より、

$0.67 = (\text{個々のデータ} - 150.3) / 5 \rightarrow \text{個々のデータ} = 153.35 \dots$ 控え

(2) 「内容量の多い方から 5% の重さ」というのは、右側の紫色の部分の面積 = 0.05

これを満たす z の値を求めればよい。

これを満たす z の値は、正規分布表より、 $z = 1.64$ となる。

$z = (\text{個々のデータ} - 150.3) / 5$ より、

$1.64 = (\text{個々のデータ} - 150.3) / 5 \rightarrow \text{個々のデータ} = 158.5$

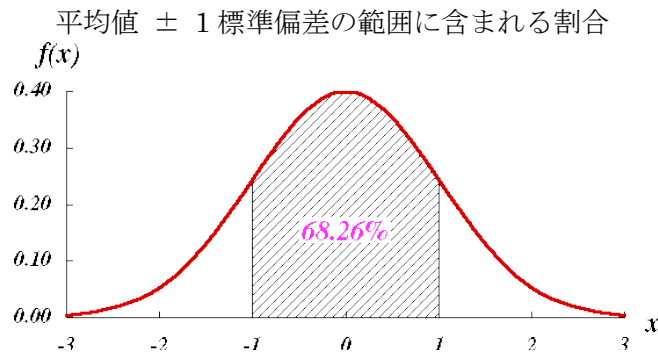
¹⁵ これは「N (0, 1²)」と表記することが多い。

標準正規分布と標準偏差 (σ)

確率変数 X が $N(\mu, \sigma^2)$ に従う時、平均 μ からのずれが $\pm 1\sigma$ 以下の範囲に X が含まれる確率は 68.26%, $\pm 2\sigma$ 以下だと 95.44%, さらに $\pm 3\sigma$ だと 99.74% となる。

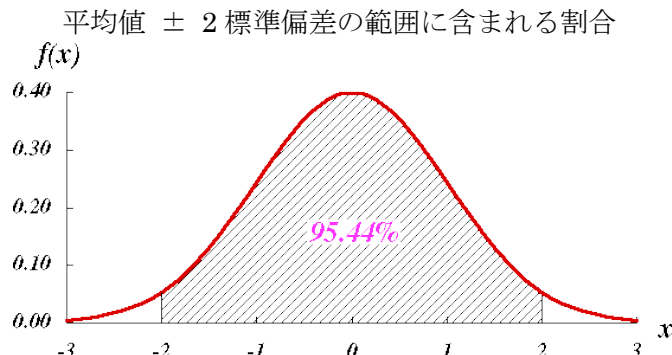
標準正規分布表で $z=1$ (1 標準偏差 = 1σ) のとき、確率が 0.3413。

➔ 平均値 (= 0) を中心として $\pm 1\sigma$ の値をとるのは全体の 約 68.26% ($34.13\% \times 2$)。
分布が正規分布であることが条件



標準正規分布表で $z=2$ (2 標準偏差 = 2σ) のとき、確率は 0.4772 である。

➔ 平均値 (= 0) を中心として $\pm 2\sigma$ の値をとるのは全体の約 95.44% ($47.72\% \times 2$)。
分布が正規分布であることが条件

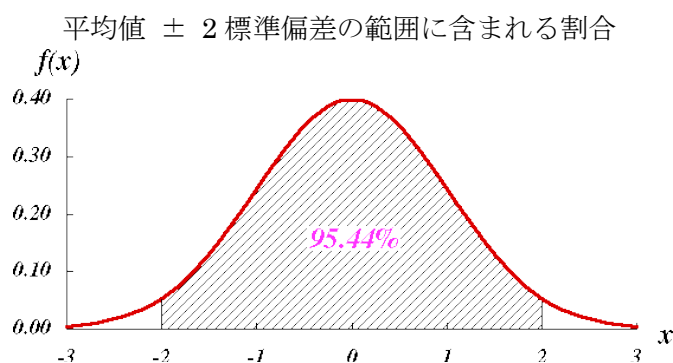


同様に $z=3$ (3 標準偏差 = 3σ) のとき、確率が 0.4987 である。

➔ 平均値 (= 0) を中心として $\pm 3\sigma$ の値をとるのは全体の 99.74% ($49.87\% \times 2$)。
分布が正規分布であることが条件

標準正規分布表で $z=2$ (2 標準偏差 = 2σ) のとき、確率は 0.4772 である。

➔ 平均値 (= 0) を中心として $\pm 2\sigma$ の値をとるのは全体の約 95.44% ($47.72\% \times 2$)。
分布が正規分布であることが条件



統計学では、的中確率をできるだけ 95% ぴったりにとろうとする
余分な 0.44 を取り除く必要がある
[- 2 以上 + 2 以下] の区間を若干狭める (標準正規確率表の面積 **0.4750** 参照)

「標準正規分布の 95% の予言的中区間」

「- 1.96 以上 + 1.96 以下」

$$z = (x - \mu) \div \sigma \quad \text{ゆえ}$$

データ x が、平均値が μ で SD が σ の正規分布に従う場合の 95% 予言的中区間は、

$$- 1.96 \leq (x - \mu) \div \sigma \leq + 1.96$$

という不等式を解いて得られる範囲である

あなたが 100 枚のコインを同時に投げるとします。その際、出るコインの表の枚数を予言するとき、95% 予言的中になる範囲を求めなさい。

解法：

	表	裏
1 回目	50	50
2 回目	49	41
3 回目	53	47
...
98 回目	55	45
99 回目	42	48
100 回目	51	49

「N 枚のコイン投げで出る表の枚数」は近似的に「平均が $N/2$ で SD が $\sqrt{N/2}$ の一般正規分布になる」

「100 枚のコインを同時に投げた時に出る表の枚数」を多数回繰り返し観測して相対度数のヒストグラムを作成すると

➔ 「平均値が $100/2 = 50$ で、SD が $\sqrt{100/2} = 5$ の一般正規分布」のヒストグラムそっくりのものができ
このとき、95% 予言的中になる範囲は、

$$-1.96 \leq (x - \mu) \div \sigma \leq +1.96$$

$$-1.96 \leq (x - 50) \div 5 \leq +1.96$$

$$-1.96 \times 5 \leq (x - 50) \leq +1.96 \times 5$$

$$-1.96 \times 5 + 50 \leq x \leq +1.96 \times 5 + 50$$

$$40.2 \leq x \leq 59.8$$

偏差値

全体の分布を、平均値を 50、標準偏差が 10 になるように変換した値

自分の点数が全体のどの位置にあるのかを知るためには、自分の点数と全体の平均値・標準偏差を比較する必要がある

例) 偏差値が 60 ということは・・・

「平均値 ± 1 標準偏差」に全体の 68% が含まれるわけだから、
40 点 ($50-10$) と 60 点 ($50+10$) の間に全体の 68% が含まれる

40 点以下の人と 60 点以上の人を合計すると、全体の 32% ($100-68$)

分布のベルカーブは左右対称だから、60 以上の方は 16% ($=32/2$)

つまり、上位 16% に入っているという意味

エクササイズ

Q1:大学入試の模擬試験を受験したところ、偏差値が 70 であることがわかった。自分は全受験者の上位何パーセントに入っているか答えなさい。

Q2: 早稲田大学政治経済学部を 20,000 人が受験したところ、その成績は、平均値 65 点、標準偏差 10 点の正規分布に従った。

- (1)ある受験生が、75 点以上 85 点以下である確率を求めなさい。
- (2)この入学試験において、上位 10 %に入るためには、何点以上あればよいか。
- (3)この入学試験において、上位 1000 人が合格する。合格するためには何点以上必要か。