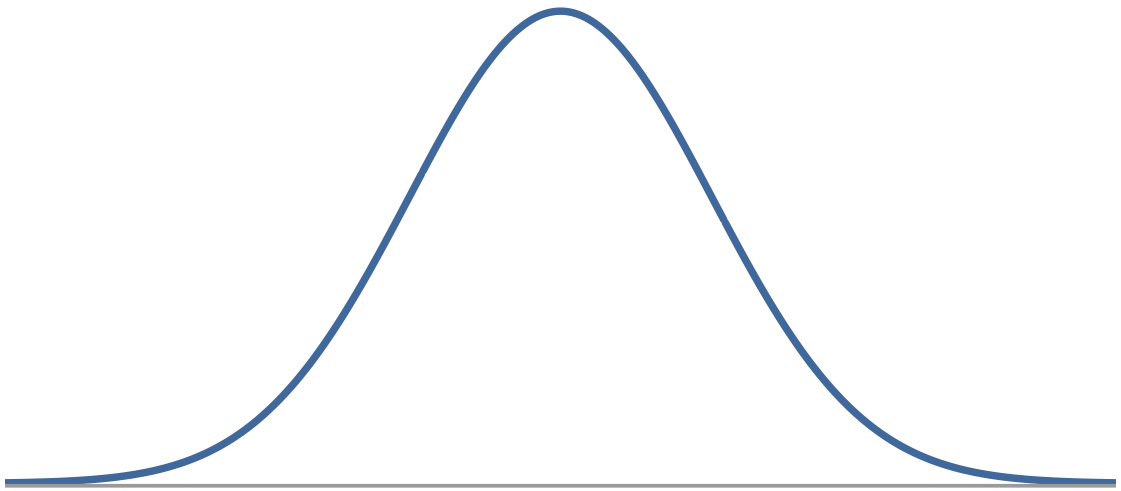
第 4 章：確率変数の分布

OpenIntro Statistics 第 4 版（日本語版）

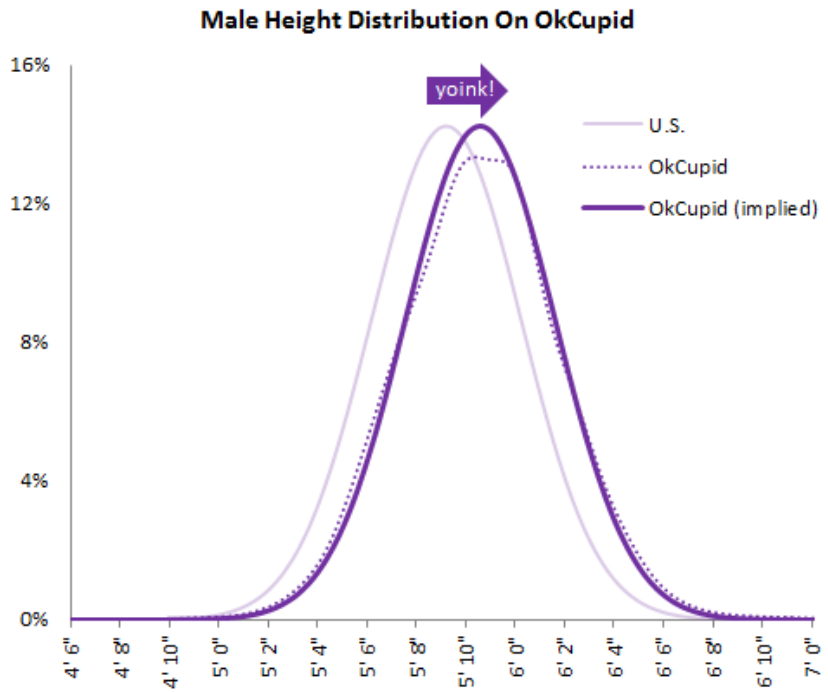
原著スライド：Mine Çetinkaya-Rundel（OpenIntro）

正規分布

- 単峰で対称な、ベル型の曲線
- 多くの変数はほぼ正規分布に従うが、完全に正規分布に従うものはない
- $N(\mu, \sigma)$ と表す $\rightarrow$ 平均 μ、標準偏差 σ の正規分布



男性の身長

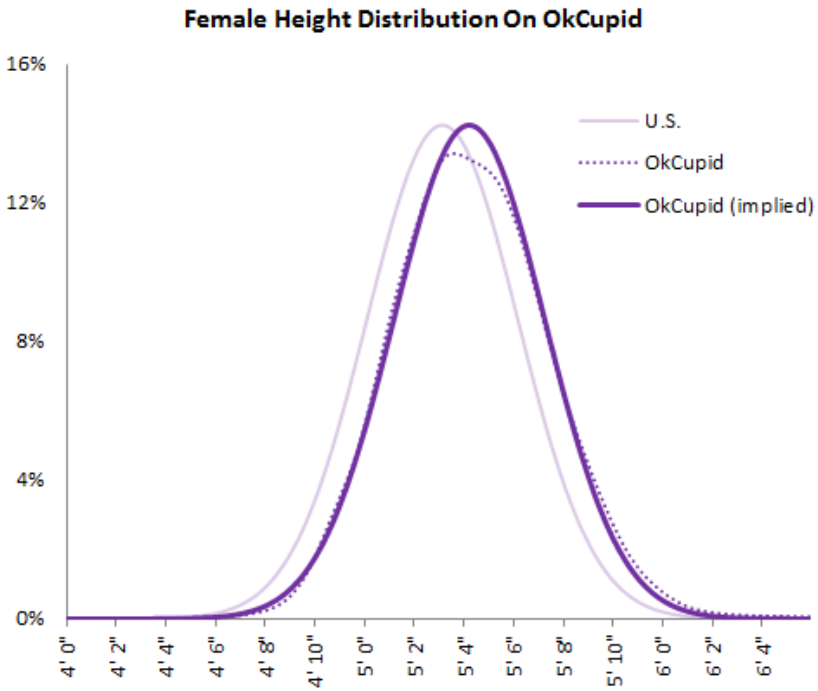


「OkCupid の男性の身長は、期待される正規分布にほぼ従っているが、全体的に右にずれている。男性はほぼ普遍的に数センチ多めに申告する傾向がある。」

「約 5 フィート 8 インチ付近から、点線の曲線の上部がさらに右に傾いていることがわかる。これは、6 フィートに近づくにつれて男性が少し多めに切り上げていることを意味し、その羨望の心理的な基準値を目指している。」

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

女性の身長



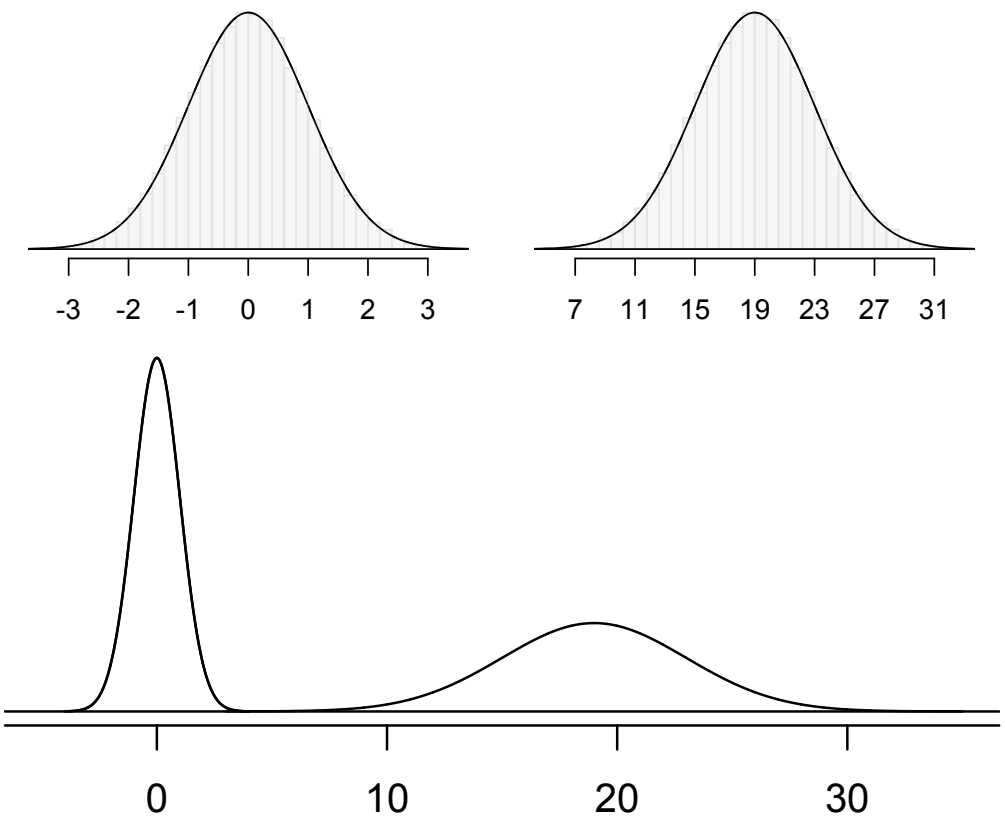
「女性のデータを調べたところ、身長の誇張は同様に広まっていたが、特定の基準値に向けた急激な変化はなかった。」

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

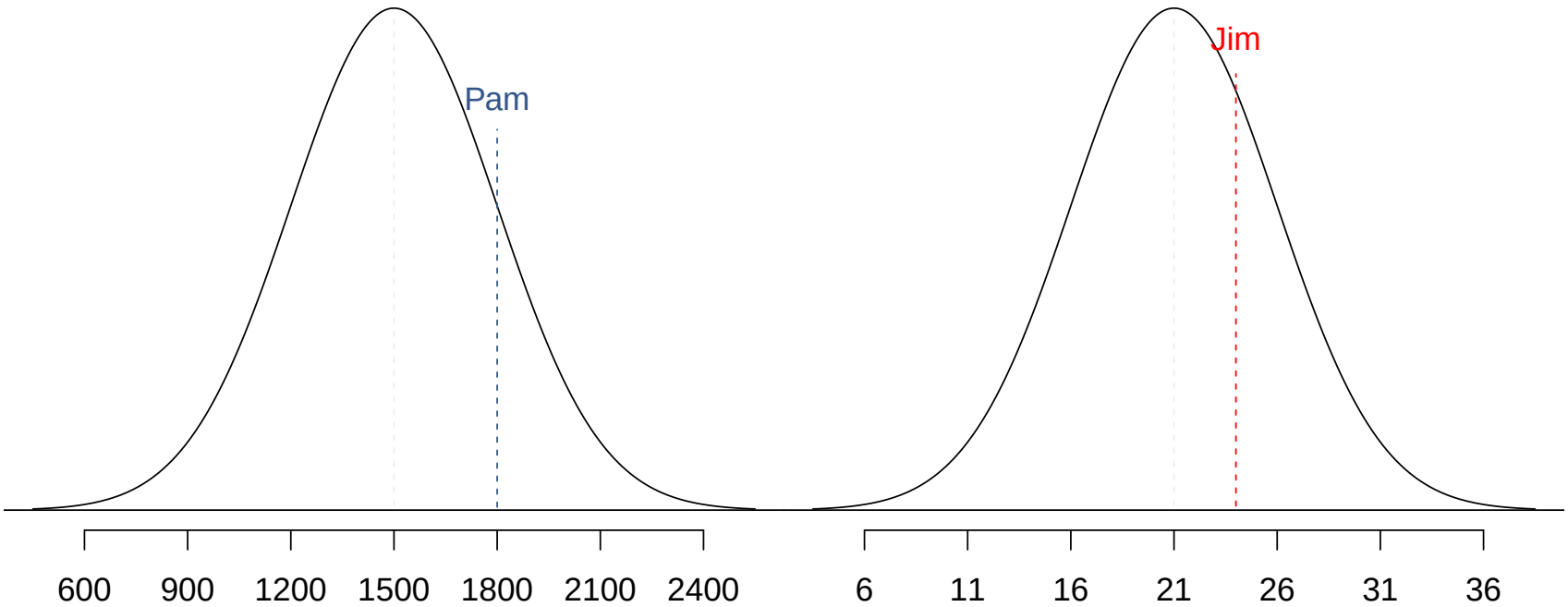
パラメータの異なる正規分布

μ：平均、σ：標準偏差

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



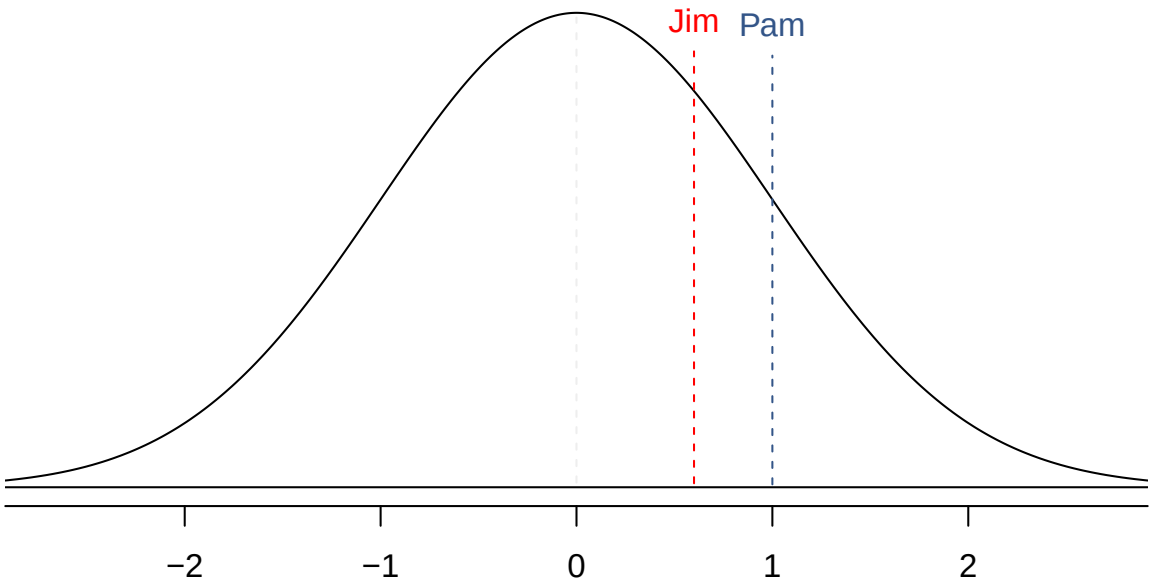
SAT のスコアは平均 1500、標準偏差 300 の正規分布にほぼ従う。ACT のスコアは平均 21、標準偏差 5 の正規分布にほぼ従う。大学の入学担当者が 2 人の志願者のうち、他の受験者に対してどちらが標準化テストでより良い成績を収めたかを判断したい。SAT で 1800 点を獲得したパムと、ACT で 24 点を取ったジムのどちらが優れているか？



Z スコアによる標準化

これら 2 つの生の点数を直接比較することはできないため、代わりに各観測値が平均から何標準偏差離れているかを比較する。

- パムのスコアは平均より $\frac{1800-1500}{300} = 1$ 標準偏差上にある。
- ジムのスコアは平均より $\frac{24-21}{5} = 0.6$ 標準偏差上にある。



Z スコアによる標準化（続き）

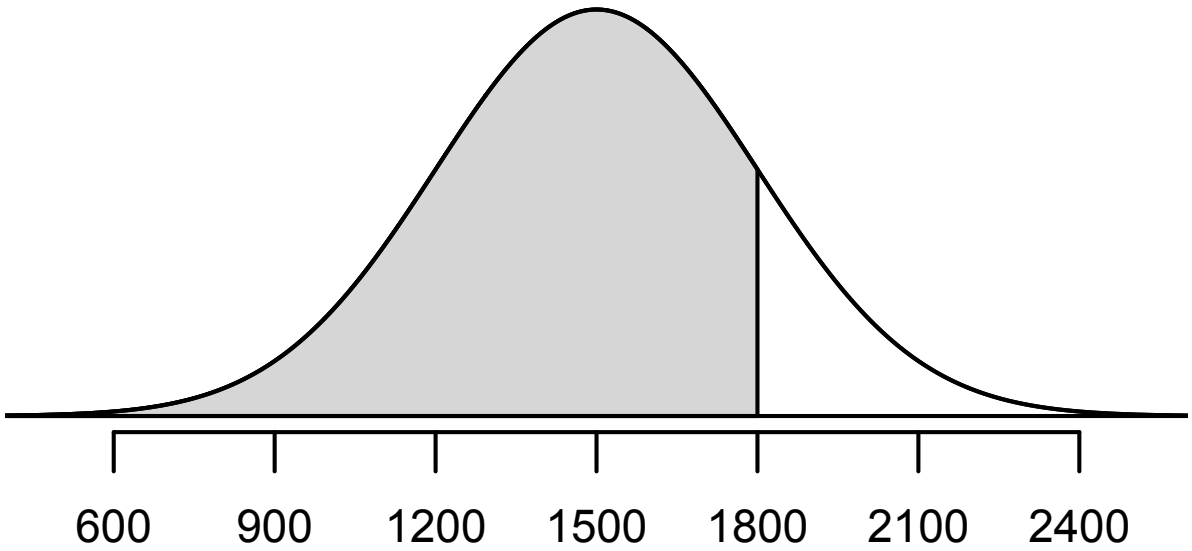
- これらは標準化スコア、または Z スコアと呼ばれる。
- 観測値の Z スコアは、それが平均から何標準偏差上または下にあるかを示す。

$$Z = \frac{\text{観測値} - \text{平均}}{SD}$$

- Z スコアはどんな形の分布に対しても定義できるが、分布が正規分布の場合にのみ Z スコアを使ってパーセンタイルを計算できる。
- 平均から 2SD 以上離れている観測値（$|Z| > 2$）は通常、異常値と見なされる。

パーセンタイル

- パーセンタイルとは、ある特定のデータ点より下に位置する観測値の割合のことである。
- グラフ的には、パーセンタイルはその観測値の左側の確率分布曲線の下の面積である。



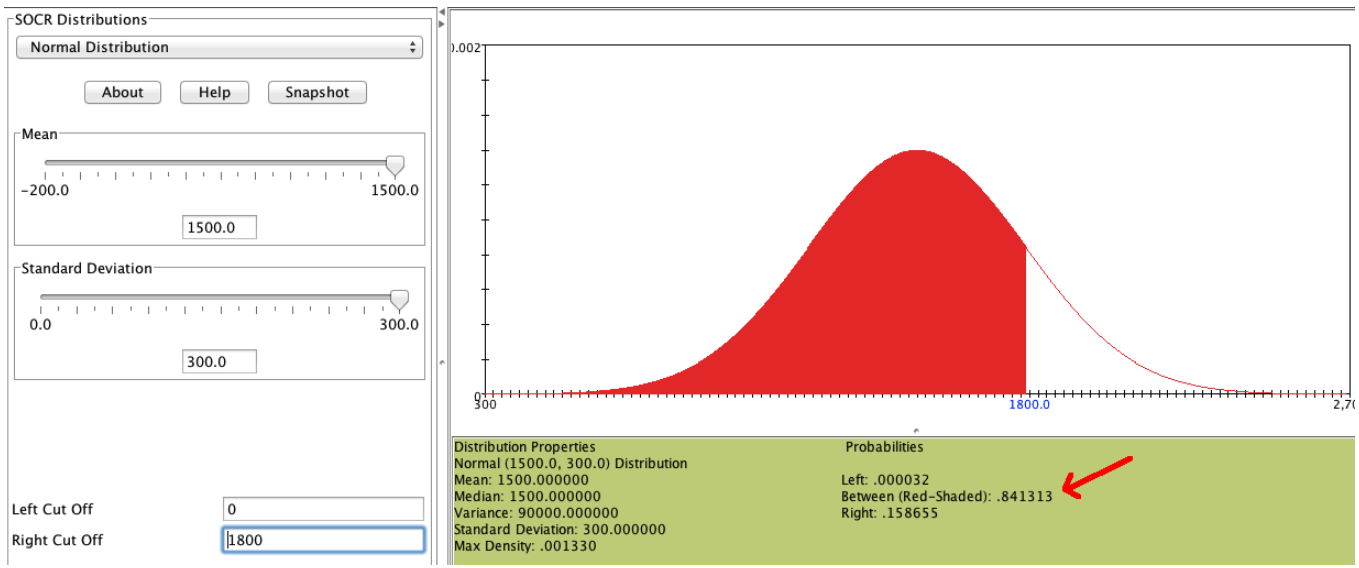
パーセンタイルの計算 — 計算機を使う

曲線の下の面積（パーセンタイル）を計算するには様々な方法がある：

- R：

```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

- アプレット：*https://gallery.shinyapps.io/dist_calc/*



パーセンタイルの計算 — 表を使う

	Z の小数第 2 位									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

シックスシグマ

「シックスシグマプロセスという言葉は、グラフに示すように、プロセス平均と最も近い規格限界との間に 6 つの標準偏差がある場合、事実上すべての製品が規格を満たすという考え方に由来する。」

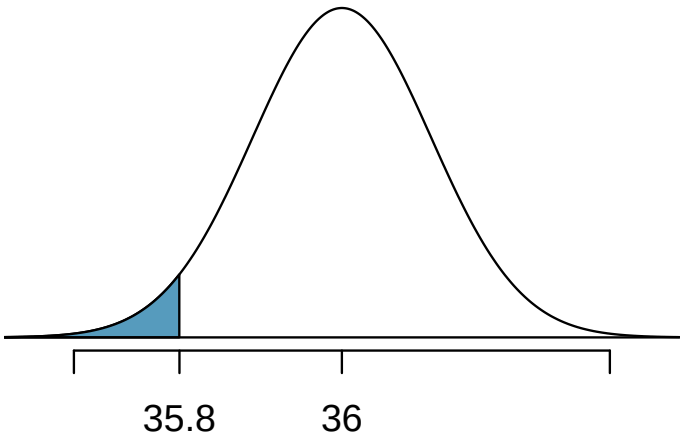
http://en.wikipedia.org/wiki/Six_Sigma

品質管理

ハインツのケチャップ工場では、ケチャップのボトルに入る量が平均 36 オンス、標準偏差 0.11 オンスの正規分布に従うとされている。30 分ごとに生産ラインからボトルが 1 本選ばれ、その内容量が正確に記録される。ケチャップの量が 35.8 オンス未満または 36.2 オンス超の場合、そのボトルは品質管理検査に不合格となる。ケチャップが 35.8 オンス未満のボトルは何パーセントか？

$X =$ ボトル内のケチャップの量とする：

$X \sim N(\mu = 36, \sigma = 0.11)$



$$Z = \frac{35.8 - 36}{0.11} = -1.82$$

正確な確率の求め方 — R を使う

```
> pnorm(-1.82, mean = 0, sd = 1)
[1] 0.0344
```

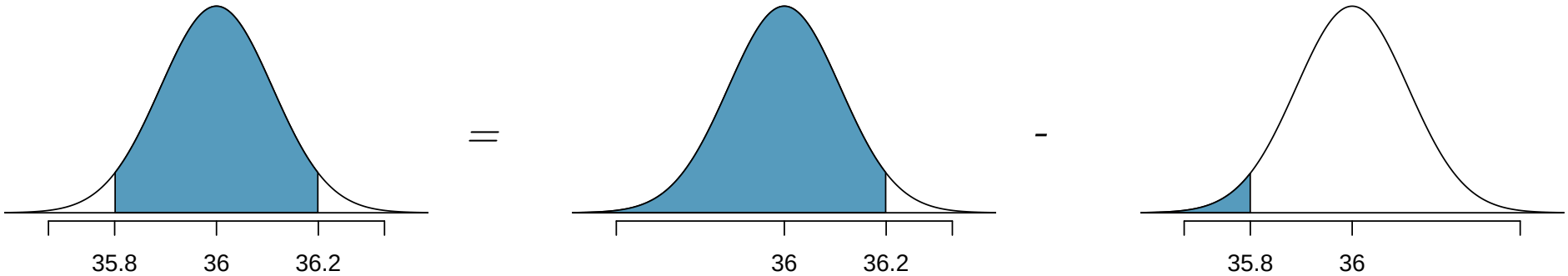
または

```
> pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.0345
```

練習問題

品質管理検査に合格するボトルは何パーセントか？

(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%

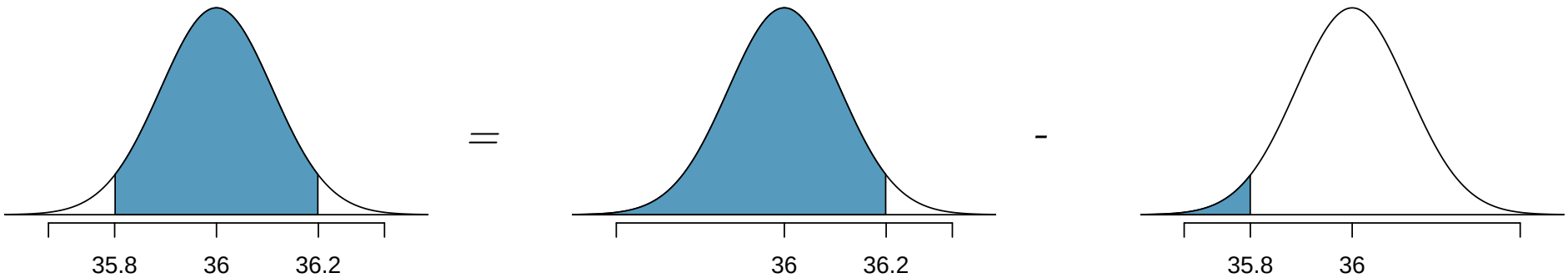


$$
\begin{aligned}
Z_{35.8} &= \frac{35.8-36}{0.11} = -1.82 \\
Z_{36.2} &= \frac{36.2-36}{0.11} = 1.82 \\
P(35.8 < X < 36.2) &= P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312
\end{aligned}
$$

練習問題

品質管理検査に合格するボトルは何パーセントか？

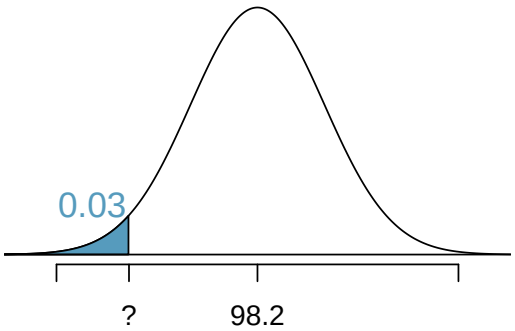
(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) *93.12%*
(e) 96.56%



$$
\begin{aligned}
Z_{35.8} &= \frac{35.8-36}{0.11} = -1.82 \\
Z_{36.2} &= \frac{36.2-36}{0.11} = 1.82 \\
P(35.8 < X < 36.2) &= P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312
\end{aligned}
$$

カットオフ点の求め方

健康な人間の体温は平均 98.2°F、標準偏差 0.73°F の正規分布にほぼ従う。人間の体温の下位 3%のカットオフは何 °F か？



$$
\begin{aligned}
P(X < x) &= 0.03 \\
&\rightarrow P(Z < \text{-1.88}) = 0.03 \\
Z &= \frac{\text{観測値} - \text{平均}}{SD} \\
&\rightarrow \frac{x - 98.2}{0.73} = -1.88 \\
x &= (-1.88 \times 0.73) + 98.2 = 96.8^\circ F
\end{aligned}
$$

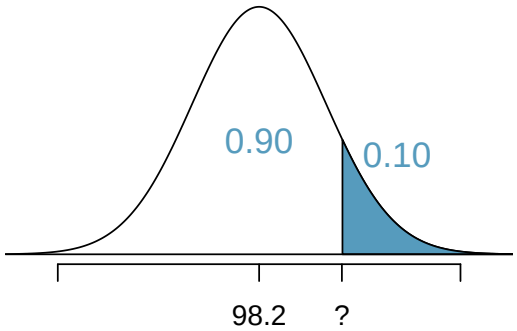
```
> qnorm(0.03)
[1] -1.880794
```

Mackowiak, Wasserman, and Levine (1992), *A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick*.

練習問題

健康な人間の体温は平均 98.2°F、標準偏差 0.73°F の正規分布にほぼ従う。
人間の体温の上位 10%のカットオフは何 °F か？

(a) 97.3°F
(b) 99.1°F
(c) 99.4°F
(d) 99.6°F

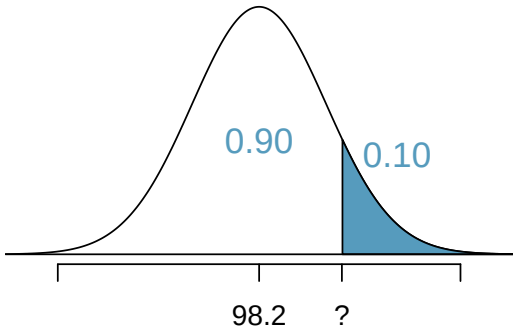


$$
\begin{aligned}
P(X > x) &= 0.10 \rightarrow P(Z < 1.28) = 0.90 \\
Z &= \frac{\text{観測値} - \text{平均}}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28 \\
x &= (1.28 \times 0.73) + 98.2 = 99.1
\end{aligned}
$$

練習問題

健康な人間の体温は平均 98.2°F、標準偏差 0.73°F の正規分布にほぼ従う。
人間の体温の上位 10%のカットオフは何 °F か？

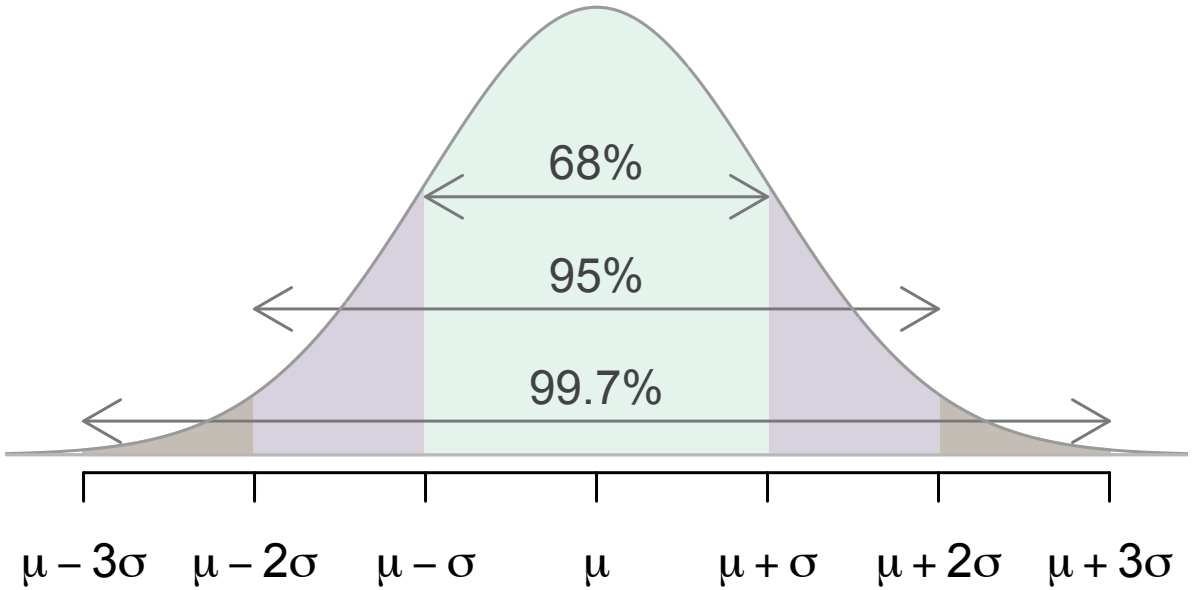
(a) 97.3°F
(b) *99.1° F*
(c) 99.4°F
(d) 99.6°F



$$
\begin{aligned}
P(X > x) &= 0.10 \to P(Z < 1.28) = 0.90 \\
Z &= \frac{\text{観測値} - \text{平均}}{SD} \to \frac{x - 98.2}{0.73} = 1.28 \\
x &= (1.28 \times 0.73) + 98.2 = 99.1
\end{aligned}
$$

68-95-99.7 則

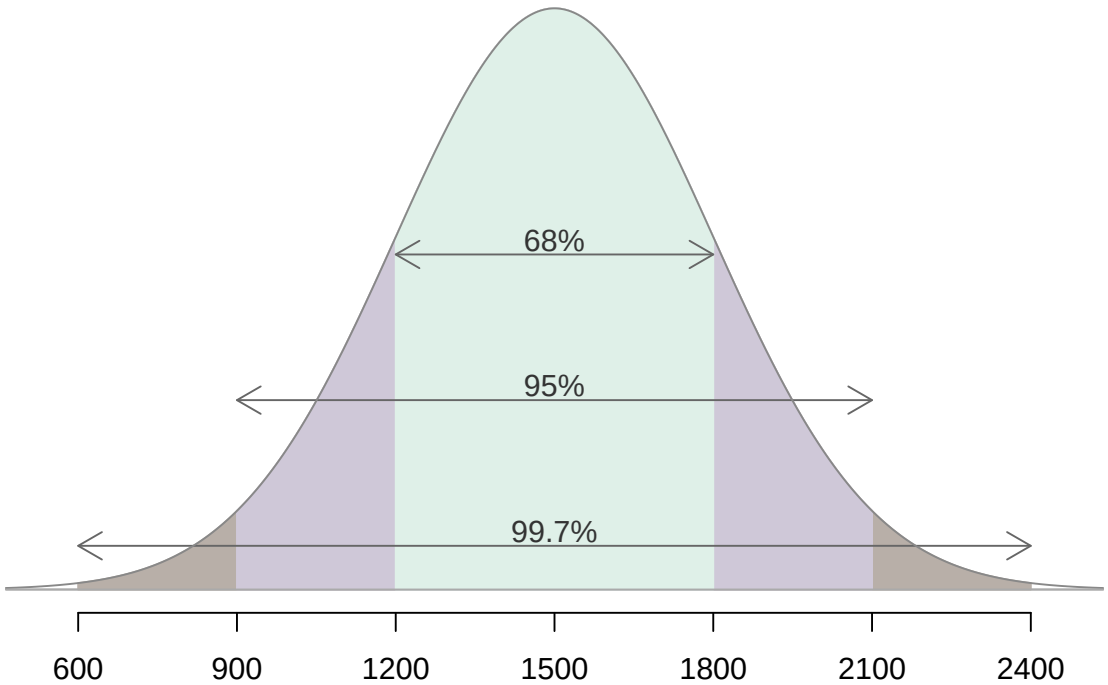
- ほぼ正規分布に従うデータでは、
  - 約 68%が平均から 1SD 以内に収まる、
  - 約 95%が平均から 2SD 以内に収まる、
  - 約 99.7%が平均から 3SD 以内に収まる。
- 平均から 4SD、5SD、またはそれ以上離れた観測値が存在することもあるが、データがほぼ正規分布に従う場合、このような出来事は非常にまれである。



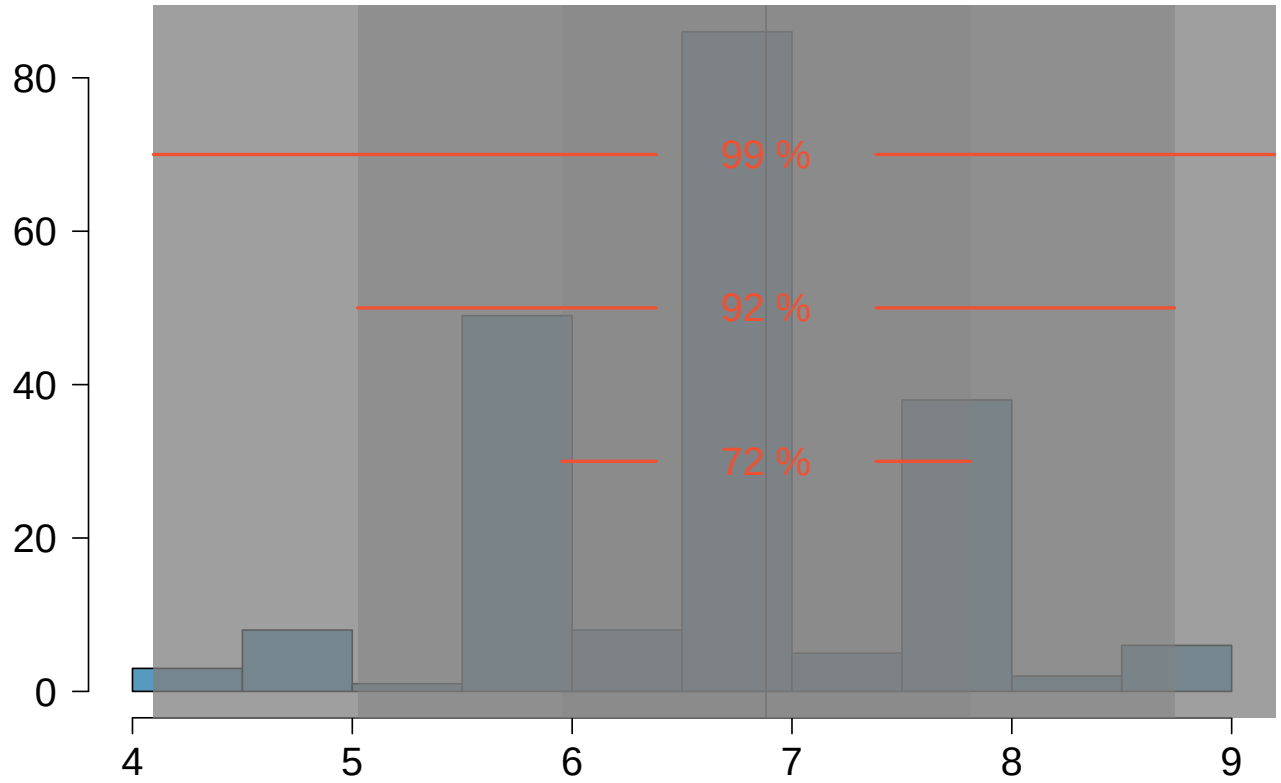
68-95-99.7 則を使った変動性の記述

SAT のスコアは平均 1500、標準偏差 300 の正規分布にほぼ従う。

- SAT で 1200〜1800 点の間に収まる学生は約 68%。
- SAT で 900〜2100 点の間に収まる学生は約 95%。
- SAT で 600〜2400 点の間に収まる学生は約 99.7%。



平日の睡眠時間



- 平均 = 6.88 時間、SD = 0.92 時間
- データの 72%は平均から 1SD 以内：6.88 ± 0.93
- データの 92%は平均から 2SD 以内：$6.88 \pm 2 \times 0.93$
- データの 99%は平均から 3SD 以内：$6.88 \pm 3 \times 0.93$

練習問題

次のうち誤りはどれか？

(a) 右に歪んだ分布では、Z スコアの大半は負である。

(b) 歪んだ分布では、平均の Z スコアが 0 と異なる場合がある。

(c) 正規分布では、IQR は $2 \times SD$ より小さい。

(d) Z スコアは、あるデータ点が分布の残りのデータと比べてどれだけ異常かを判断するのに役立つ。

練習問題

次のうち誤りはどれか？

(a) 右に歪んだ分布では、Z スコアの大半は負である。

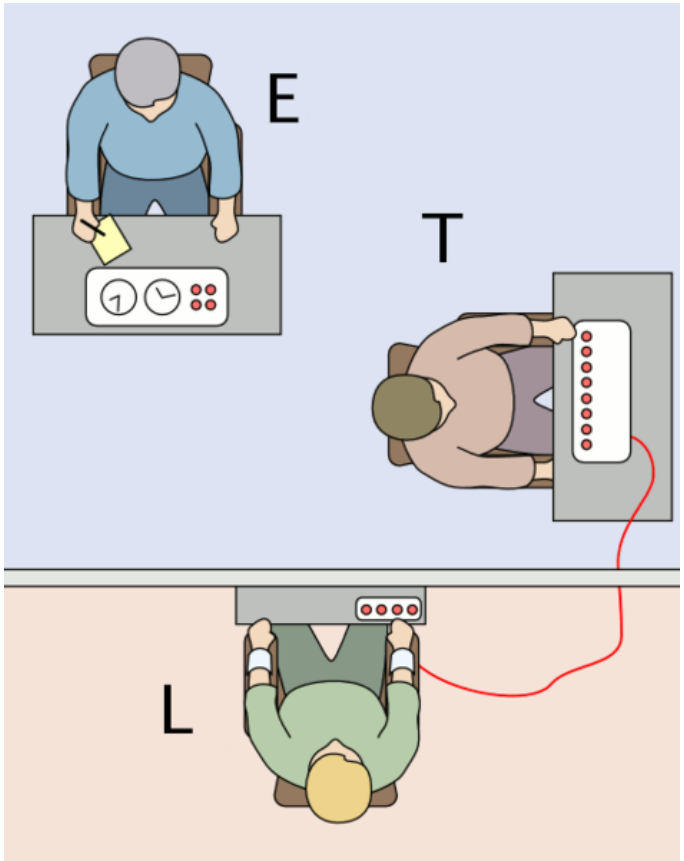
(b) 歪んだ分布では、平均の Z スコアが 0 と異なる場合がある。

(c) 正規分布では、IQR は $2 \times SD$ より小さい。

(d) Z スコアは、あるデータ点が分布の残りのデータと比べてどれだけ異常かを判断するのに役立つ。

ミルグラム実験

- イェール大学の心理学者スタンレー・ミルグラムは、1963 年から権威への服従に関する一連の実験を行った。
- 実験者（E）は教師（T）（実験の被験者）に対し、学習者（L）が質問に誤答するたびに強烈な電気ショックを与えるよう命令する。
- 学習者は実際には俳優であり、電気ショックは本物ではないが、教師がショックを与えるたびに事前録音された音が再生される。



http://en.wikipedia.org/wiki/File:Milgram_Experiment_v2.png

ミルグラム実験（続き）

- これらの実験は、個人の良心に反する行為を指示する権威者に対して、研究参加者がどれほど従うかを測定した。
- ミルグラムは、約 65%の人が権威に従い、そのようなショックを与えることを発見した。
- 長年にわたる研究により、この数字はコミュニティや時代を超えてほぼ一定であることが示唆されている。

ベルヌーイ確率変数

- ミルグラムの実験における各人は試行と見なすことができる。
- 強烈なショックを与えることを拒否した場合は成功、ショックを与えた場合は失敗とラベルを付ける。
- ショックを与えることを拒否した人はわずか 35%であるため、成功確率は $p = 0.35$ である。
- 個々の試行に 2 つの結果しかない場合、それをベルヌーイ確率変数という。

幾何分布

スミス博士はミルグラムの実験を繰り返したいが、強烈なショックを与えない人が見つかるまでサンプリングしたいと考えている。最初の人でやめる確率は何か？

$$P(\text{1人目が拒否}) = 0.35$$

……3 人目の人でやめる確率は？

$$\begin{aligned} P(\text{1 人目と 2 人目がショック、3 人目が拒否}) &= \frac{S}{0.65} \times \frac{S}{0.65} \times \frac{R}{0.35} \\ &= 0.65^2 \times 0.35 \approx 0.15 \end{aligned}$$

……10 人目の人でやめる確率は？

$$\begin{aligned} P(\textit{9}\text{ 人がショック、}\textit{10}\text{ 人目が拒否}) &= \underbrace{\frac{S}{0.65} \times \cdots \times \frac{S}{0.65}}_{\text{9個}} \times \frac{R}{0.35} \\ &= 0.65^9 \times 0.35 \approx 0.0072 \end{aligned}$$

幾何分布（続き）

幾何分布は、独立同一分布（*iid*）のベルヌーイ確率変数において、最初の成功までの待ち時間を記述する。

- 独立性：各試行の結果は互いに影響しない
- 同一性：成功確率は各試行で同じ

幾何確率

成功確率を p、失敗確率を $(1-p)$、独立試行回数を n とすると

$$P(n\text{回目の試行で初めて成功}) = (1-p)^{n-1}p$$

サイコロを 6 回振って初めて 6 の目が出る確率を幾何分布で計算できるか？ 成功（6 の目が出る）と失敗（6 の目が出ない）は明確に定義されており、各試行でどちらかが必ず起こることに注意せよ。

(a) いいえ、サイコロを振ると 2 つ以上の結果がある

(b) はい、計算できる

$$P(6\text{回目で初めて}\ 6\ \text{の目}) = \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right) \approx 0.067$$

サイコロを 6 回振って初めて 6 の目が出る確率を幾何分布で計算できるか？ 成功（6 の目が出る）と失敗（6 の目が出ない）は明確に定義されており、各試行でどちらかが必ず起こることに注意せよ。

(a) いいえ、サイコロを振ると 2 つ以上の結果がある

(b) はい、計算できる

$$P(6\text{回目で初めて}\ 6\ \text{の目}) = \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right) \approx 0.067$$

期待値

スミス博士は、ショックを与えることを拒否する最初の人を見つけるまで、何人の人を試験すると予想されるか？

幾何分布の期待値、すなわち平均は $\frac{1}{p}$ と定義される。

$$\mu = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

スミス博士は、ショックを与えることを拒否する最初の人を見つけるまで、2.86 人の人を試験すると予想される。
しかし、整数でない人数をどのように試験するのか？

期待値とその変動性

幾何分布の平均と標準偏差

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

- スミス博士の実験に戻ると：

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.35}{0.35^2}} = 2.3$$

- スミス博士は、ショックを与えることを拒否する最初の人を見つけるまで、平均 2.86 人（誤差 ±2.3 人）の人を試験すると予想される。
- これらの値は、実験を非常に多くの回数繰り返す文脈でのみ意味をなす。

実験に参加するために 4 人を無作為に選んだとする。そのうちちょうど 1 人がショックを与えることを拒否する確率は何か？

これらの人をアレン（A）、ブリタニー（B）、キャロライン（C）、デイミアン（D）と呼ぼう。以下の 4 つのシナリオそれぞれが「ちょうど 1 人が拒否する」という条件を満たす：

シナリオ 1: $\underset{\text{(A) 拒否}}{0.35} \times \underset{\text{(B) ショック}}{0.65} \times \underset{\text{(C) ショック}}{0.65} \times \underset{\text{(D) ショック}}{0.65} = 0.0961$

シナリオ 2: $\underset{\text{(A) ショック}}{0.65} \times \underset{\text{(B) 拒否}}{0.35} \times \underset{\text{(C) ショック}}{0.65} \times \underset{\text{(D) ショック}}{0.65} = 0.0961$

シナリオ 3: $\underset{\text{(A) ショック}}{0.65} \times \underset{\text{(B) ショック}}{0.65} \times \underset{\text{(C) 拒否}}{0.35} \times \underset{\text{(D) ショック}}{0.65} = 0.0961$

シナリオ 4: $\underset{\text{(A) ショック}}{0.65} \times \underset{\text{(B) ショック}}{0.65} \times \underset{\text{(C) ショック}}{0.65} \times \underset{\text{(D) 拒否}}{0.35} = 0.0961$

4 人のうちちょうど 1 人がショックを与えることを拒否する確率は、これらすべての確率の和である。

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

二項分布

前のスライドの問いは、n 回の試行中の成功回数 k の確率（$n = 4$ 回中 $k = 1$ 回の成功）を求めており、この確率を次のように計算した

$$\text{シナリオ数} \times P(1\text{ つのシナリオ})$$

- シナリオ数：これを計算するより簡単な方法がある。後ほど説明する……
- $P(1\text{ つのシナリオ}) = p^k\ (1-p)^{(n-k)}$

  成功確率の成功回数乗、失敗確率の失敗回数乗

二項分布は、成功確率 p の n 回の独立なベルヌーイ試行においてちょうど k 回成功する確率を記述する。

シナリオ数の数え方

先ほど、ちょうど 1 人がショックを与えることを拒否するという条件に合うすべてのシナリオを書き出した。もし n がより大きく、k が 1 以外の場合、例えば $n = 9$、$k = 2$ では：

*RR*SSSSSSS
S*RR*SSSSSS
SS*RR*SSSSS
...
SS*R*SS*R*SSS
...
SSSSSSS*RR*

すべてのシナリオを書き出すことは非常に面倒で誤りが起きやすい。

シナリオ数の計算

二項係数

二項係数（choose 関数）は、n 回の試行で k 回成功する方法の数を計算するのに役立つ。

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $k = 1$、$n = 4$: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$
- $k = 2$、$n = 9$: $\binom{9}{2} = \frac{9!}{2!(9-2)!} = \frac{9\times8\times7!}{2\times1\times7!} = \frac{72}{2} = 36$

注：*R* を使っても計算できる：

```
> choose(9,2)
[1] 36
```

二項係数の性質

次のうち誤りはどれか？

(a) n 回の試行で 1 回成功する方法は n 通りある、$\binom{n}{1} = n$。

(b) n 回の試行で n 回成功する方法は 1 通りしかない、$\binom{n}{n} = 1$。

(c) n 回の試行で n 回失敗する方法は 1 通りしかない、$\binom{n}{0} = 1$。

(d) n 回の試行で $n-1$ 回成功する方法は $n-1$ 通りある、$\binom{n}{n-1} = n-1$。

二項係数の性質

次のうち誤りはどれか？

(a) n 回の試行で 1 回成功する方法は n 通りある、$\binom{n}{1} = n$。

(b) n 回の試行で n 回成功する方法は 1 通りしかない、$\binom{n}{n} = 1$。

(c) n 回の試行で n 回失敗する方法は 1 通りしかない、$\binom{n}{0} = 1$。

(d) n 回の試行で $n-1$ 回成功する方法は $n-1$ 通りある、$\binom{n}{n-1} = n-1$。

二項分布（続き）

二項確率

p が成功確率、$(1-p)$ が失敗確率、n が独立試行回数、k が成功回数を表すとき

$$P(n\text{回中 } k\text{回の成功}) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

二項分布を適用するために満たす必要のない条件はどれか？

(a) 試行は独立でなければならない

(b) 試行回数 n は固定されていなければならない

(c) 各試行の結果は成功または失敗に分類されなければならない

(d) 望ましい成功回数 k は試行回数より大きくなければならない

(e) 成功確率 p は各試行で同じでなければならない

二項分布を適用するために満たす必要のない条件はどれか？

(a) 試行は独立でなければならない

(b) 試行回数 n は固定されていなければならない

(c) 各試行の結果は成功または失敗に分類されなければならない

(d) 望ましい成功回数 k は試行回数より大きくなければならない

(e) 成功確率 p は各試行で同じでなければならない

2012 年のギャラップ調査によると、アメリカ人の 26.2%が肥満である。10 人のアメリカ人の無作為標本の中で、ちょうど 8 人が肥満である確率は何か？

(a) かなり高い

(b) かなり低い

Gallup: *http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx*, January 23, 2013.

2012 年のギャラップ調査によると、アメリカ人の 26.2%が肥満である。10 人のアメリカ人の無作為標本の中で、ちょうど 8 人が肥満である確率は何か？

(a) かなり高い

(b) かなり低い

Gallup: *http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx*, January 23, 2013.

2012 年のギャラップ調査によると、アメリカ人の 26.2%が肥満である。10 人のアメリカ人の無作為標本の中で、ちょうど 8 人が肥満である確率は何か？

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

2012 年のギャラップ調査によると、アメリカ人の 26.2%が肥満である。10 人のアメリカ人の無作為標本の中で、ちょうど 8 人が肥満である確率は何か？

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2 \ = 45 \times 0.262^8 \times 0.738^2 = 0.0005$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

誕生日問題

無作為に選んだ 2 人が同じ誕生日を持つ確率は何か？

かなり低く、$\frac{1}{365} \approx 0.0027$ である。

366 人の中で、少なくとも 2 人が同じ誕生日を持つ確率は何か？

確実に 1 である！（閏年の誕生日の可能性を除く。）

誕生日問題（続き）

121 人の中で、少なくとも 2 人（1 組のペア）が同じ誕生日を持つ確率は何か？

計算はやや複雑だが、121 人の中で一致がない確率の余事象として考えることができる。

$$
\begin{aligned}
P(\text{一致なし}) &= 1 \times \left(1 - \tfrac{1}{365}\right) \times \left(1 - \tfrac{2}{365}\right) \\
&\quad \times \cdots \times \left(1 - \tfrac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}} \approx 0 \\
P(\text{少なくとも 1 組のペア}) &\approx 1
\end{aligned}
$$

期待値

2012 年のギャラップ調査によると、アメリカ人の 26.2%が肥満である。

100 人のアメリカ人の無作為標本の中で、何人が肥満であると予想されるか？

- 簡単に計算すると、$100 \times 0.262 = 26.2$ である。
- より形式的には、$\mu = np = 100 \times 0.262 = 26.2$ となる。
- しかし、これはすべての 100 人の無作為標本でちょうど 26.2 人が肥満であることを意味するわけではない。実際、それは不可能である。標本によってこの値は小さかったり大きかったりする。この値がどれだけ変動すると予想されるか？

期待値とその変動性

二項分布の平均と標準偏差

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

- 肥満率に戻ると：

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

- 無作為に選んだ 100 人のアメリカ人のうち 26.2 人が肥満であると予想され、標準偏差は 4.4 である。

注：二項分布の平均と標準偏差が常に整数であるとは限らないが、それで構わない。これらの値は平均的に期待されることを表している。

異常な観測値

平均から 2 標準偏差以上離れている観測値は異常と見なすという考えと、先ほど計算した平均と標準偏差を使って、100 人の無作為標本における肥満者数の妥当な範囲を計算できる。

$$26.2 \pm (2 \times 4.4) = (17.4, 35)$$

2012 年 8 月のギャラップ調査によると、アメリカ人の 13%がホームスクーリングは子どもに優れた教育を提供すると考えている。1,000 人のアメリカ人の無作為標本でこの意見を持つ人が 100 人のみだった場合、それは異常と見なされるか？

(a) いいえ

(b) はい

$$\mu = np = 1,000 \times 0.13 = 130$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

方法 1: 通常の観測値の範囲：$130 \pm 2 \times 10.6 = (108.8, 151.2)$
100 はこの範囲外なので、異常と見なされる。

方法 2: 観測値の Z スコア：$Z = \frac{x - \text{平均}}{SD} = \frac{100-130}{10.6} = -2.83$
100 は平均より *2SD* 以上下にあるため、異常と見なされる。

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

2012 年 8 月のギャラップ調査によると、アメリカ人の 13%がホームスクーリングは子どもに優れた教育を提供すると考えている。1,000 人のアメリカ人の無作為標本でこの意見を持つ人が 100 人のみだった場合、それは異常と見なされるか？

(a) いいえ

(b) はい

$$\mu = np = 1,000 \times 0.13 = 130$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

方法 1: 通常の観測値の範囲：$130 \pm 2 \times 10.6 = (108.8, 151.2)$
100 はこの範囲外なので、異常と見なされる。

方法 2: 観測値の Z スコア：$Z = \frac{x - \text{平均}}{SD} = \frac{100-130}{10.6} = -2.83$
100 は平均より *2SD* 以上下にあるため、異常と見なされる。

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

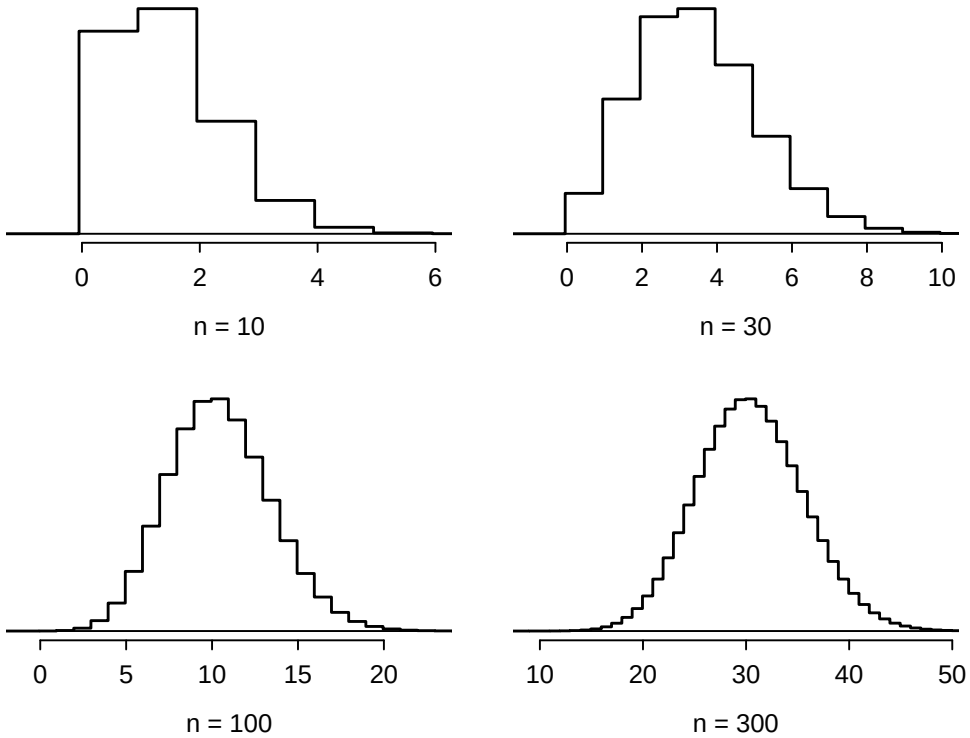
二項分布の形状

この活動にはウェブアプレットを使用する。
https://gallery.shinyapps.io/dist_calc/ にアクセスし、左のドロップダウンメニューで「Binomial coin experiment」を選ぶ。

- 試行回数を 20、成功確率を 0.15 に設定する。成功回数の分布の形状を説明せよ。
- p を 0.15 に固定したまま、成功回数の分布が単峰かつ対称になるのに必要な最小の標本サイズを求めよ。
- さらに検討：
  - n を一定に保ちながら p が変化すると分布の形状はどうなるか？
  - p を一定に保ちながら n が変化すると分布の形状はどうなるか？

成功回数の分布

$p = 0.10$、$n = 10$、30、100、300 の二項モデルから得られたサンプルのヒストグラム。n が増加するとどうなるか？



十分大きいとはどれくらいか？

期待される成功回数と失敗回数がともに少なくとも 10 であれば、標本サイズは十分大きいと見なされる。

$$np \geq 10 \qquad \text{かつ} \qquad n(1-p) \geq 10$$

$10 \times 0.13 = 1.3; 10 \times (1 - 0.13) = 8.7$

以下の 4 組の二項分布パラメータのうち、正規分布で近似できる分布はどれか？

(a) $n = 100, p = 0.95$

(b) $n = 25, p = 0.45$

(c) $n = 150, p = 0.05$

(d) $n = 500, p = 0.015$

以下の 4 組の二項分布パラメータのうち、正規分布で近似できる分布はどれか？

(a) $n = 100, p = 0.95$

(b) $n = 25, p = 0.45 \quad \rightarrow 25 \times 0.45 = 11.25; 25 \times 0.55 = 13.75$

(c) $n = 150, p = 0.05$

(d) $n = 500, p = 0.015$

Facebook ユーザーの分析

最近の研究によると「Facebook ユーザーは与えるよりも多くを受け取っている」。例えば：

- 標本内の Facebook ユーザーの 40%が友達リクエストを送ったが、63%が少なくとも 1 件のリクエストを受け取った
- ユーザーは友達のコンテンツに平均 14 回「いいね」を押したが、自分のコンテンツには平均 20 回「いいね」がついた
- ユーザーは 9 件の個人メッセージを送ったが、12 件を受け取った
- ユーザーの 12%が写真で友達をタグ付けしたが、35%が自分自身を写真でタグ付けされた

このパターンをどのように説明できるか？

パワーユーザーが一般ユーザーよりはるかに多くのコンテンツを投稿している。

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

この研究では、Facebook ユーザーの約 25%がパワーユーザーと見なされることもわかった。同じ研究で、平均的な Facebook ユーザーは 245 人の友達を持つことが明らかになった。245 人の友達を持つ平均的な Facebook ユーザーが、パワーユーザーと見なされる友達が 70 人以上いる確率は何か？ 必要な仮定を述べよ。

$n = 245$、$p = 0.25$ が与えられ、確率 $P(K \geq 70)$ を求める。独立性が必要であり、ここでは仮定する（Facebook のデータにアクセスできれば確認できる）。

$$\begin{aligned} P(X \geq 70) &= P(K = 70 \text{ または } K = 71 \text{ または } \cdots \text{ または } K = 245) \\ &= P(K = 70) + P(K = 71) + \cdots + P(K = 245) \end{aligned}$$

これは膨大な計算になりそうだ……

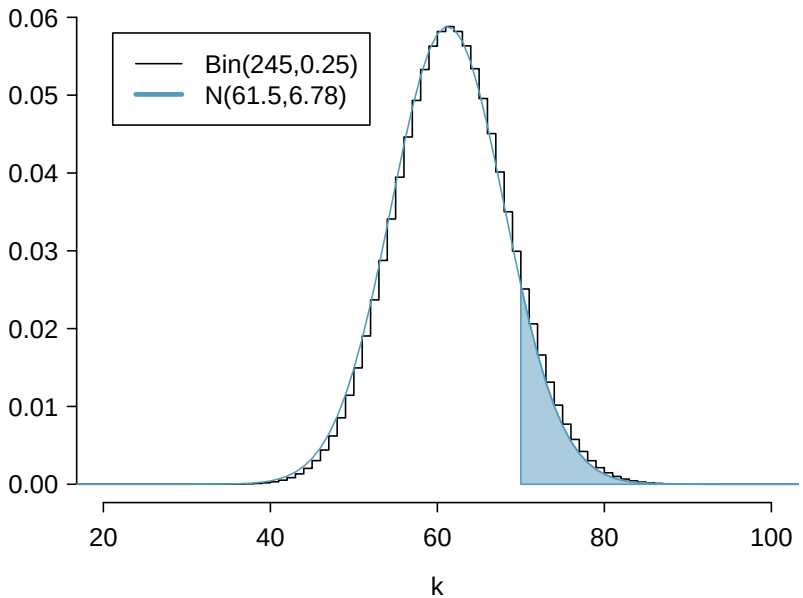
二項分布の正規近似

標本サイズが十分大きい場合、パラメータ n と p の二項分布は、パラメータ $\mu = np$、$\sigma = \sqrt{np(1-p)}$ の正規モデルで近似できる。

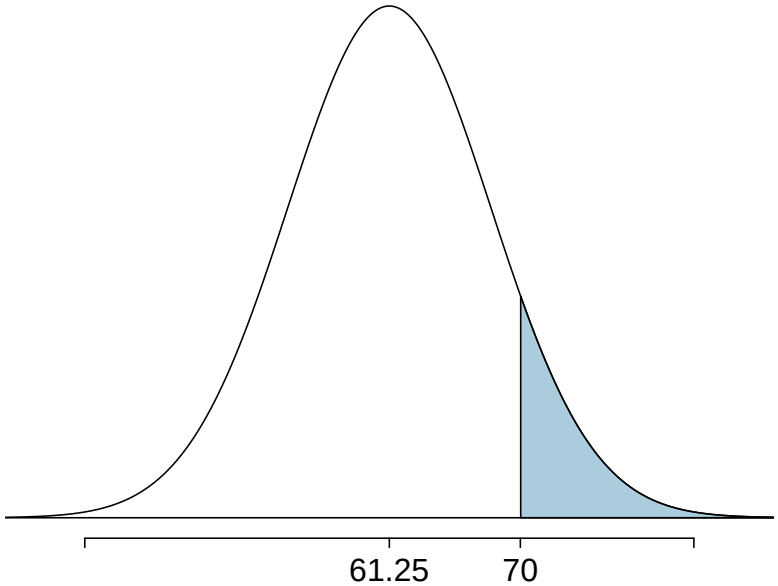
- Facebook のパワーユーザーの場合、$n = 245$、$p = 0.25$ である。

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$。



245 人の友達を持つ平均的な Facebook ユーザーが、パワーユーザーと見なされる友達が 70 人以上いる確率は何か？



$$Z = \frac{\text{観測値} - \text{平均}}{SD}$$
$$= \frac{70 - 61.25}{6.78} = 1.29$$

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$

```
> pnorm(1.29)
[1] 0.9014747
```

小さい区間での正規近似の限界

- 二項分布に対する正規近似は、条件が満たされていても、少数の値の範囲の確率を推定する際に精度が低下する傾向がある。
- 値の区間に対するこの近似は、カットオフ値を両方向に 0.5 ずつ拡張することで通常改善される。
- 正規近似を適用する際にこの修正を加えるヒントは、観測値の範囲を調べる際に最も役立つ。裾確率を計算する際にもこの修正を適用することは可能だが、区間が通常かなり広いため修正の利点は消えることが多い。

負の二項分布

- 負の二項分布は、n 回目の試行で k 回目の成功が起こる確率を記述する。
- 負の二項分布の状況を識別するのに役立つ 4 つの条件：
  1. 試行は独立である。
  2. 各試行の結果は成功または失敗に分類される。
  3. 成功確率（p）は各試行で同じである。
  4. 最後の試行は成功でなければならない。

  最初の 3 つの条件は二項分布と共通している。

負の二項分布

$\mathsf{P}(n$ 回目の試行で k 回目の成功$) = \binom{n-1}{k-1}\, p^k\, (1-p)^{n-k}$、
ここで p は個々の試行が成功する確率である。すべての試行は独立と仮定する。

心理学研究室でアルバイトをしている大学生が、研究に参加するカップルを 10 組募集するよう頼まれた。彼女は学生センターの前に立ち、建物を出る 5 人おきに 1 人に対して、交際中かどうか、もし交際中であれば重要な他者と一緒に研究に参加したいかを尋ねることにした。このような人を見つける確率が 10%だとする。目標を達成するまでに 30 人に尋ねる必要がある確率は何か？

与えられた情報：$p = 0.10$、$k = 10$、$n = 30$。30 回目の試行で 10 回目の成功の確率を求めるため、負の二項分布を使用する。

$$
\begin{aligned}
P(\text{30 回目で 10 回目の成功}) &= \binom{29}{9} \times 0.10^{10} \times 0.90^{20} \\
&= 10,015,005 \times 0.10^{10} \times 0.90^{20} \\
&= 0.00012
\end{aligned}
$$

二項分布 vs. 負の二項分布

負の二項分布は二項分布とどのように異なるか？

- 二項分布の場合、通常は固定された試行回数があり、代わりに成功の回数を考える。
- 負の二項分布の場合、固定された回数の成功を観測するまでに何回の試行が必要かを調べ、最後の観測が成功であることが必要である。

練習問題

次のうち、目的の確率を計算するために負の二項分布を使う状況はどれか？

(a) 5 歳の男の子の身長が 42 インチより高い確率。
(b) 10 回のソフトボール投球のうち 3 回成功する確率。
(c) ポーカーでストレートフラッシュの手が配られる確率。
(d) 最初のヒットの前に 8 回ミスをする確率。
(e) 8 回目の試行で 3 回目のボールに当たる確率。

練習問題

次のうち、目的の確率を計算するために負の二項分布を使う状況はどれか？

(a) 5 歳の男の子の身長が 42 インチより高い確率。

(b) 10 回のソフトボール投球のうち 3 回成功する確率。

(c) ポーカーでストレートフラッシュの手が配られる確率。

(d) 最初のヒットの前に 8 回ミスをする確率。

(e) *8* 回目の試行で *3* 回目のボールに当たる確率。

ポアソン分布

- ポアソン分布は、固定された母集団において個人が独立である場合、短い単位時間あたりの大きな母集団における稀な事象の数を推定するのに役立つことが多い。
- ポアソン分布のレートとは、主に固定された母集団において単位時間あたりに起こる事象の平均回数のことであり、通常 λ で表される。
- レートを使って、単位時間内に稀な事象がちょうど k 回観測される確率を記述できる。

ポアソン分布

P(稀な事象を k 回観測) $= \frac{\lambda^k e^{-\lambda}}{k!}$、
ここで k は 0、1、2、……の値をとり、$k!$ は k の階乗を表す。
$e \approx 2.718$ は自然対数の底である。
この分布の平均と標準偏差はそれぞれ λ と $\sqrt{\lambda}$ である。

ある発展途上国の農村地域では、停電がポアソン分布に従い、週平均 2 回発生するとする。ある週に停電がちょうど 1 回だけ発生する確率を計算せよ。

$\lambda = 2$ が与えられる。

$$
\begin{aligned}
P(\text{週に 1 回だけ停電}) &= \frac{2^1 \times e^{-2}}{1!} \\
&= \frac{2 \times e^{-2}}{1} \\
&= 0.27
\end{aligned}
$$

ある発展途上国の農村地域では、停電がポアソン分布に従い、週平均 2 回発生するとする。ある特定の日に停電が 3 回発生する確率を計算せよ。

週単位の停電レートが与えられているが、この問いに答えるには、まずある特定の日の平均停電レートを計算する必要がある：$\lambda_{日} = \frac{2}{7} = 0.2857$。停電の確率は曜日によらず同じ、すなわち独立であると仮定することに注意する。

$$
\begin{aligned}
P(\text{ある日に 3 回の停電}) &= \frac{0.2857^3 \times e^{-0.2857}}{3!} \\
&= \frac{0.2857^3 \times e^{-0.2857}}{6} \\
&= 0.0029
\end{aligned}
$$

ポアソン分布は適用できるか？

- 考慮中の事象が稀であり、母集団が大きく、事象が互いに独立に発生する場合、確率変数はポアソン分布に従う可能性がある
- しかし、事象が実際には独立でない状況も考えられる。例えば、ある夏の間の結婚式の数の確率に関心がある場合、週末の方が結婚式に人気があることを考慮すべきである。
- この場合、異なる時間帯に異なるレートを持たせることができれば、ポアソンモデルが合理的な場合もある。たとえば、週末のレートを平日より高くモデル化できる。
- ポアソン分布のレートを別の変数（曜日）に対してモデル化するという考えは、一般化線形モデルと呼ばれるより高度な手法の基礎を形成する。本講義の範囲を超えるが、第 7 章と第 8 章で線形モデルの基礎を議論する。

練習問題

次のどの分布に従う確率変数が正の整数以外の値をとることができるか？

(a) ポアソン分布
(b) 負の二項分布
(c) 二項分布
(d) 正規分布
(e) 幾何分布

練習問題

次のどの分布に従う確率変数が正の整数以外の値をとることができるか？

(a) ポアソン分布
(b) 負の二項分布
(c) 二項分布
(d) 正規分布
(e) 幾何分布