第3章：確率

OpenIntro Statistics 第4版（日本語版）

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

確率的プロセス

- 確率的プロセスとは、何が起こりうるかはわかっているが、実際にどの結果が起こるかわからない状況のことである。
- 例：コイン投げ、サイコロの目、iTunes シャッフル、明日の株式市場の上下など。
- 実際には確率的でなくても、プロセスを確率的にモデル化することが有用な場合がある。

MP3 Players › Stories › iTunes: Just how random is random?

iTunes: Just how random is random?

By David Braue on 08 March 2007

- Introduction
- Say You, Say What?
- A role for labels?
- The new random

Think that song has appeared in your playlists just a few too many times? David Braue puts the randomness of Apple's song shuffling to the test -- and finds some surprising results.

Quick -- think of a number between one and 20. Now think of another one, and another, and another.

Starting to repeat yourself? No surprise: in practice, many series of random numbers are far less random than you would think.

Computers have the same problem. Although all systems are able to pick random numbers, the method they use is often tied to specific other numbers -- for example, the time -- that means you could get a very similar series of 'random' numbers in different situations.

This tendency manifests itself in many ways. For anyone who uses their iPod heavily, you've probably noticed that your supposedly random 'shuffling' iPod seems to be particularly fond of the Bee Gees, Melissa Etheridge or Pavarotti. Look at a random playlist that iTunes generates for you, and you're likely to notice several songs from one or two artists, while other artists go completely unrepresented.

http://www.cnet.com.au/

itunes-just-how-random-is-random-339274094.htm

確率

- 確率にはいくつかの解釈があるが、確率が従うべき数学的規則についてはほぼ完全に一致している。
  - $P(A) =$ 事象 A の確率
  - $0 \leq P(A) \leq 1$
- 頻度論的解釈：
  - ある結果の確率は、確率的プロセスを無限回観察したときにその結果が起こる割合である。
- ベイズ的解釈：
  - ベイズ主義者は確率を主観的な信念の度合いとして解釈する：同じ事象でも、異なる人が異なる確率を割り当てることがある。
  - 過去 20 年間の計算技術と方法の革命的進歩によって広まった。

練習問題

次のどの事象が最も驚くべきものか？

(a) 10 回コインを投げて、ちょうど 3 回表が出る
(b) 100 回コインを投げて、ちょうど 3 回表が出る
(c) 1000 回コインを投げて、ちょうど 3 回表が出る

練習問題

次のどの事象が最も驚くべきものか？

(a) 10 回コインを投げて、ちょうど 3 回表が出る
(b) 100 回コインを投げて、ちょうど 3 回表が出る
(c) *1000 回コインを投げて、ちょうど 3 回表が出る*

大数の法則

大数の法則とは、観測数が増えるにつれて、ある特定の結果が起こる割合 $\hat{p}_n$ が、その結果の確率 p に収束することを述べたものである。

大数の法則（続き）

公正なコインを投げたとき、最初の 10 回すべてで表が出た場合、次の投げで表が出る確率はいくつだと思うか？ 0.5、0.5 より小さい、それとも 0.5 より大きい？

$$\underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{H}\ \underline{?}$$

- 確率はやはり 0.5 であり、次の投げで表が出る確率は 50% である。

$$P(H \text{ on } 11^{th} \text{ toss}) = P(T \text{ on } 11^{th} \text{ toss}) = 0.5$$

- コインは「裏を出すべき」ではない。
- 大数の法則についての一般的な誤解は、確率的プロセスは過去に起こったことを補正するはずだというものであるが、これは誤りであり、ギャンブラーの誤謬（または平均の法則）とも呼ばれる。

互いに排反な結果とそうでない結果

互いに排反（*mutually exclusive*）な結果：同時には起こり得ない。

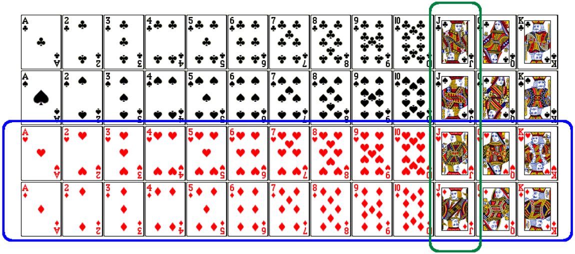
- 1 回のコイン投げの結果が表と裏の両方になることはない。
- 学生が同じ科目で合格と不合格の両方になることはない。
- 1 枚のカードがエースとクイーンの両方であることはない。

互いに排反でない結果：同時に起こり得る。

- 学生が同じ学期に統計学で A、経済学で A をとることがある。

互いに排反でない事象の和事象

よくシャッフルされた１組のトランプから、ジャックまたは赤いカードを引く確率は何か？



$$P(\text{ジャックまたは赤}) = P(\text{ジャック}) + P(\text{赤}) - P(\text{ジャックかつ赤})$$

$$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

Figure from *http://www.milefoot.com/math/discrete/counting/cardfreq.htm*.

練習問題

無作為に選ばれた学生が、大麻を合法化すべきと考えているか、または親の政治的見解に同意している確率は何か？

	親の政治観に同意		
大麻合法化	なし	あり	合計
反対	11	40	51
賛成	36	78	114
合計	47	118	165

(a) $\frac{40+36-78}{165}$

(b) $\frac{114+118-78}{165}$

(c) $\frac{78}{165}$

(d) $\frac{78}{188}$

(e) $\frac{11}{47}$

練習問題

無作為に選ばれた学生が、大麻を合法化すべきと考えているか、または親の政治的見解に同意している確率は何か？

	親の政治観に同意		
大麻合法化	なし	あり	合計
反対	11	40	51
賛成	36	78	114
合計	47	118	165

(a) $\frac{40+36-78}{165}$

(b) $\frac{114+118-78}{165}$

(c) $\frac{78}{165}$

(d) $\frac{78}{188}$

(e) $\frac{11}{47}$

まとめ

一般加法定則

$$P(A\ \text{または}\ B) = P(A) + P(B) - P(A\ \text{かつ}\ B)$$

注：互いに排反な事象では $P(A\ \text{かつ}\ B) = 0$ なので、上式は $P(A\ \text{または}\ B) = P(A) + P(B)$ に簡略化される。

確率分布

確率分布は、起こりうるすべての事象とそれらが起こる確率を一覧にしたものである。

- 1 人の子どもの性別の確率分布：

事象	男	女
確率	0.5	0.5

- 確率分布の規則：
  1. 列挙された事象は互いに排反でなければならない
  2. 各確率は 0 以上 1 以下でなければならない
  3. 確率の合計は 1 でなければならない
- 2 人の子どもの性別の確率分布：

事象	MM	FF	MF	FM
確率	0.25	0.25	0.25	0.25

練習問題

調査で、回答者の 52%が民主党員だと答えた。このサンプルから無作為に選ばれた回答者が共和党員である確率はいくつか？

(a) 0.48
(b) 0.48 より大きい
(c) 0.48 より小さい
(d) 与えられた情報だけでは計算できない

練習問題

調査で、回答者の 52%が民主党員だと答えた。このサンプルから無作為に選ばれた回答者が共和党員である確率はいくつか？

(a) 0.48
(b) 0.48 より大きい
(c) 0.48 より小さい
(d) 与えられた情報だけでは計算できない

唯一の *2* つの政党が共和党と民主党である場合は *(a)* が可能である。しかし、政党に所属しない人やこれら *2* つ以外の政党に所属する人もいる可能性があるため、*(c)* も可能である。*(b)* は合計確率が *1* を超えることになるため、絶対に不可能である。

標本空間と余事象

標本空間は、試行で起こりうるすべての結果の集合である。

- ある夫婦に 1 人の子どもがいる場合、その性別の標本空間は？ $S = \{M, F\}$
- ある夫婦に 2 人の子どもがいる場合、その性別の標本空間は？ $S = \{MM, FF, FM, MF\}$

余事象は、確率の和が 1 になる 2 つの互いに排反な事象である。

- ある夫婦に 1 人の子どもがいる。その子が男の子でないとわかった場合、その性別は？{ ~~M~~, F } → 男の子と女の子は余事象である。
- ある夫婦に 2 人の子どもがいる。両方が女の子でないとわかった場合、可能な性別の組み合わせは？{ MM, ~~FF~~, FM, MF }

独立

2 つのプロセスが独立であるとは、一方の結果を知っても他方の結果についての有用な情報が得られない場合をいう。

- 最初の投げでコインが表になったことを知っても、2 回目の投げでコインがどちらになるかについての有用な情報は得られない。$\rightarrow$ 2 回のコイン投げの結果は独立である。
- 最初に引いたカードがエースであることを知ると、2 回目の引きでエースを引く確率についての有用な情報が得られる。$\rightarrow$ デッキからの 2 回の引き（非復元）の結果は従属である。

練習問題

2013 年 1 月 9 日〜12 日、SurveyUSA がノースカロライナ州の居住者 500 人の無作為標本を対象に、銃の広範な所有は法を守る市民を犯罪から守るか、社会をより危険にするかを質問した。回答者全体の 58%が「市民を守る」と回答した。白人回答者の 67%、黒人回答者の 28%、ヒスパニック回答者の 64%がこの見解を共有していた。次のうち正しいものはどれか？

銃所有に関する意見と人種・民族は最も可能性が高いのは

(a) 余事象
(b) 互いに排反
(c) 独立
(d) 従属
(e) 互いに排反

http://www.surveyusa.com/client/PollReport.aspx?g=a5f460ef-bba9-484b-8579-1101ea26421b

練習問題

2013 年 1 月 9 日〜12 日、SurveyUSA がノースカロライナ州の居住者 500 人の無作為標本を対象に、銃の広範な所有は法を守る市民を犯罪から守るか、社会をより危険にするかを質問した。回答者全体の 58%が「市民を守る」と回答した。白人回答者の 67%、黒人回答者の 28%、ヒスパニック回答者の 64%がこの見解を共有していた。次のうち正しいものはどれか？

銃所有に関する意見と人種・民族は最も可能性が高いのは

(a) 余事象
(b) 互いに排反
(c) 独立
(d) 従属
(e) 互いに排反

http://www.surveyusa.com/client/PollReport.aspx?g=a5f460ef-bba9-484b-8579-1101ea26421b

独立性の確認

P(B が真であるとき A が起こる) $= P(A \mid B) = P(A)$ ならば、A と B は独立である。

*P(*市民を守る*) = 0.58*

*P(*無作為に選ばれたノースカロライナ州居住者が銃所有は市民を守ると答える，その居住者が白人である場合*) =*
*P(*市民を守る | 白人*) = 0.67*

*P(*市民を守る | 黒人*) = 0.28*

*P(*市民を守る | ヒスパニック*) = 0.64*

*P(*市民を守る*)* は人種・民族によって異なるため、銃所有に関する意見と人種・民族は最も可能性が高いのは従属である。

標本データに基づく従属性の判定

- 標本データから計算された条件付き確率が 2 つの変数間の従属性を示している場合、次のステップは、確率の観測された差が偶然に起こる可能性があるかを判定する仮説検定を行うことである。
- 条件付き確率間の観測された差が大きいほど、その差が真であるという証拠が強くなる。
- 標本が大きい場合、小さな差でも真の差の強い証拠となりうる。

P(市民を守る | 白人) = 0.67、P(市民を守る | ヒスパニック) = 0.64 であった。白人とヒスパニックで銃の広範な所有が市民を守ると思う割合に真の差があると確信するのはどちらの条件下か？ $n = 500$ か $n = 50,000$ か？

$n = 50,000$

独立事象の積の法則

$$P(A\ \text{かつ}\ B) = P(A) \times P(B)$$

より一般的には、$P(A_1\ \text{かつ}\ \cdots\ \text{かつ}\ A_k) = P(A_1) \times \cdots \times P(A_k)$

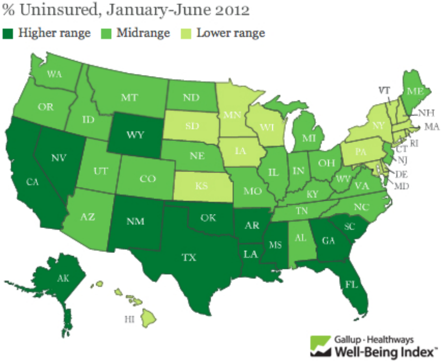
コインを 2 回投げるとき、2 回続けて裏が出る確率は何か？

$$P(\text{1 回目に裏}) \times P(\text{2 回目に裏}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

練習問題

最近のギャラップ調査によると、2012 年 6 月時点でテキサス州民の 25.5% が健康保険を持っていない。非加入率が一定であると仮定して、無作為に選んだ 2 人のテキサス州民が両方とも非加入である確率は何か？

(a) 25.5^2
(b) 0.255^2
(c) 0.255×2
(d) $(1 - 0.255)^2$

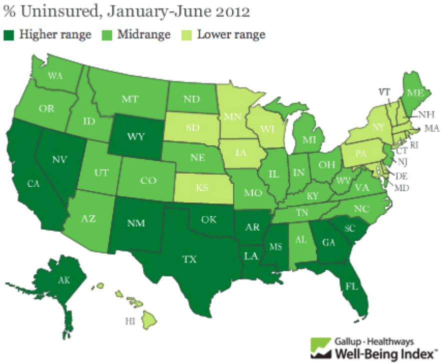


http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

練習問題

最近のギャラップ調査によると、2012 年 6 月時点でテキサス州民の 25.5% が健康保険を持っていない。非加入率が一定であると仮定して、無作為に選んだ 2 人のテキサス州民が両方とも非加入である確率は何か？

(a) 25.5^2
(b) 0.255^2
(c) 0.255×2
(d) $(1 - 0.255)^2$



http://www.gallup.com/poll/156851/uninsured-rate-stable-across-states-far-2012.aspx

互いに排反 vs. 余事象

2つの互いに排反な事象の確率の和は常に1になるか？

必ずしもそうではない。標本空間に 3 つ以上の事象がある場合がある（例：党派）。

2つの余事象の確率の和は常に1になるか？

はい、それが余事象の定義である（例：表と裏）。

すべてをまとめると…

テキサス州民を 5 人無作為に選ぶとき、少なくとも 1 人が保険未加入である確率は何か？

- テキサス州民を 5 人無作為に選んだ場合、保険未加入者の人数の標本空間は：

$$S = \{0, 1, 2, 3, 4, 5\}$$

- 少なくとも 1 人が保険未加入である場合に注目する：

$$S = \{0, 1, 2, 3, 4, 5\}$$

- 標本空間を 2 つのカテゴリに分割できる：

$$S = \{0, \text{少なくとも 1 人}\}$$

すべてをまとめると…

標本空間の確率の和は 1 でなければならないため：

$$
\begin{aligned}
P(\text{少なくとも 1 人が未加入}) &= 1 - P(\text{未加入者なし}) \\
&= 1 - [(1 - 0.255)^5] \\
&= 1 - 0.745^5 \\
&= 1 - 0.23 \\
&= 0.77
\end{aligned}
$$

少なくとも 1

$$P(\text{少なくとも 1}) = 1 - P(\text{1 つもない})$$

練習問題

ある大学の学部生の約 20%がベジタリアンまたはビーガンである。学部生 3 人の無作為標本から、少なくとも 1 人がベジタリアンまたはビーガンである確率は何か？

(a) $1 - 0.2 \times 3$

(b) $1 - 0.2^3$

(c) 0.8^3

(d) $1 - 0.8 \times 3$

(e) $1 - 0.8^3$

$$P(\text{少なくとも 1 人がベジタリアン}) =$$
$$1 - P(\text{ベジタリアンなし})$$
$$= 1 - (1 - 0.2)^3$$
$$= 1 - 0.8^3$$
$$= 1 - 0.512 = 0.488$$

練習問題

ある大学の学部生の約 20%がベジタリアンまたはビーガンである。学部生 3 人の無作為標本から、少なくとも 1 人がベジタリアンまたはビーガンである確率は何か？

(a) $1 - 0.2 \times 3$
(b) $1 - 0.2^3$
(c) 0.8^3
(d) $1 - 0.8 \times 3$
(e) $1 - 0.8^3$

$P(\text{少なくとも 1 人がベジタリアン}) =$
$1 - P(\text{ベジタリアンなし})$
$= 1 - (1 - 0.2)^3$
$= 1 - 0.8^3$
$= 1 - 0.512 = 0.488$

再発

研究者たちは、コカイン慢性使用者 72 人を 3 つのグループに無作為に割り当てた：デシプラミン（抗うつ薬）、リチウム（コカインの標準治療）、プラセボ。研究結果を以下にまとめる。

	再発あり	再発なし	合計
デシプラミン	10	14	24
リチウム	18	6	24
プラセボ	20	4	24
合計	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

患者が再発しなかった確率は何か？

	再発あり	再発なし	合計
デシプラミン	10	14	24
リチウム	18	6	24
プラセボ	20	4	24
合計	48	24	72

周辺確率

患者が再発した確率は何か？

	再発あり	再発なし	合計
デシプラミン	10	14	24
リチウム	18	6	24
プラセボ	20	4	24
合計	*48*	24	*72*

P(再発あり) $= \frac{48}{72} \approx 0.67$

結合確率

患者が抗うつ薬（デシプラミン）を投与されたかつ再発した確率は何か？

	再発あり	再発なし	合計
デシプラミン	*10*	14	24
リチウム	18	6	24
プラセボ	20	4	24
合計	48	24	*72*

$\mathrm{P}(\text{再発あり かつ デシプラミン}) = \frac{10}{72} \approx 0.14$

条件付き確率

条件付き確率

条件 B のもとで目的の結果 A が起こる条件付き確率は次のように計算される

$$P(A|B) = \frac{P(A \text{ かつ } B)}{P(B)}$$

	再発あり	再発なし	合計
デシプラミン	10	14	24
リチウム	18	6	24
プラセボ	20	4	24
合計	48	24	72

$$P(\text{再発}|\text{デシプラミン}) = \frac{P(\text{再発 かつ デシプラミン})}{P(\text{デシプラミン})} = \frac{10/72}{24/72} = \frac{10}{24} = 0.42$$

条件付き確率（続き）

患者が抗うつ薬（デシプラミン）を投与されたとわかっている場合、再発した確率は何か？

	再発あり	再発なし	合計
デシプラミン	*10*	14	*24*
リチウム	18	6	24
プラセボ	20	4	24
合計	48	24	72

$$\mathrm{P}(\text{再発} \mid \text{デシプラミン}) = \frac{10}{24} \approx 0.42$$

$$\mathrm{P}(\text{再発} \mid \text{リチウム}) = \frac{18}{24} \approx 0.75$$
$$\mathrm{P}(\text{再発} \mid \text{プラセボ}) = \frac{20}{24} \approx 0.83$$

条件付き確率（続き）

患者が再発したとわかっている場合、抗うつ薬（デシプラミン）を投与されていた確率は何か？

	再発あり	再発なし	合計
デシプラミン	*10*	14	24
リチウム	18	6	24
プラセボ	20	4	24
合計	*48*	24	72

$$\mathrm{P}(\text{デシプラミン} \mid \text{再発}) = \frac{10}{48} \approx 0.21$$

$$\mathrm{P}(\text{リチウム} \mid \text{再発}) = \frac{18}{48} \approx 0.375$$
$$\mathrm{P}(\text{プラセボ} \mid \text{再発}) = \frac{20}{48} \approx 0.42$$

一般乗法定則

- 2 つの事象が独立である場合、その結合確率は単純にそれぞれの確率の積であることを見た。事象が独立でないと考えられる場合、結合確率は少し異なる方法で計算される。
- A と B が 2 つの結果または事象を表す場合、

$$P(A \text{ かつ } B) = P(A|B) \times P(B)$$

この式は、条件付き確率の公式を変形したものに過ぎない。
- A を目的の結果、B を条件として考えると便利である。

独立性と条件付き確率

以下は、入門統計学クラスの学生の性別と専攻の（仮想の）分布である：

	社会科学	非社会科学	合計
女性	30	20	50
男性	30	20	50
合計	60	40	100

- 無作為に選ばれた学生が社会科学専攻である確率は $\frac{60}{100} = 0.6$。
- 無作為に選ばれた学生が女性であるとわかっている場合、社会科学専攻である確率は $\frac{30}{50} = 0.6$。
- $P(SS|M)$ も 0.6 に等しいため、このクラスの学生の専攻は性別に依存しない：$\mathsf{P(SS \mid F) = P(SS)}$。

独立性と条件付き確率（続き）

一般的に、$P(A|B) = P(A)$ であれば、事象 A と B は独立であるといわれる。

- 概念的に：B を与えても A について何も教えてくれない。
- 数学的に：事象 A と B が独立であれば $P(A \text{ かつ } B) = P(A) \times P(B)$ であることがわかっている。すると、

$$P(A|B) = \frac{P(A \text{ かつ } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

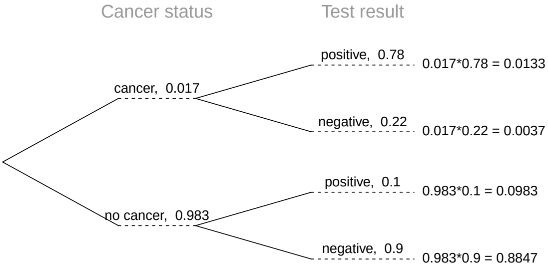
乳がん検診

- 米国がん協会の推定によれば、女性の約 1.7%が乳がんを患っている。
  http://www.cancer.org/cancer/cancerbasics/cancer-prevalence
- Susan G. Komen For The Cure Foundation によれば、マンモグラフィは実際に乳がんを患っている女性の約 78%を正確に識別する。
  http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html
- 2003 年に発表された論文によれば、すべてのマンモグラフィのうち最大 10%ががんを患っていない患者に対して偽陽性を示す。
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940

注：これらの割合は近似値であり、推定が非常に難しい。

確率の逆転

患者が乳がん検診を受けるとき、2つの競合する主張がある：患者はがんである、患者はがんでない。マンモグラフィが陽性の結果を示した場合、患者が実際にがんである確率は何か？



$$
\begin{aligned}
&P(C|+) \\
&= \frac{P(C \text{ かつ } +)}{P(+)} \\
&= \frac{0.0133}{0.0133 + 0.0983} \\
&= 0.12
\end{aligned}
$$

注：樹形図は確率を逆転させるのに有用である：$P(+|C)$ が与えられ、$P(C|+)$ を求める。

練習問題

1 回検査を受けて陽性の結果を得た女性が再検査を受けたいとする。2 回目の検査では、この特定の女性ががんである確率はいくつと仮定すべきか？

(a) 0.017
(b) 0.12
(c) 0.0133
(d) 0.88

練習問題

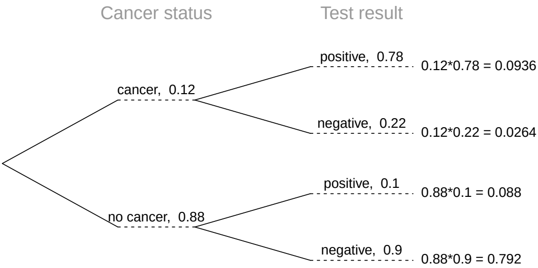
1 回検査を受けて陽性の結果を得た女性が再検査を受けたいとする。2 回目の検査では、この特定の女性ががんである確率はいくつと仮定すべきか？

(a) 0.017
(b) *0.12*
(c) 0.0133
(d) 0.88

練習問題

この 2 回目のマンモグラフィでも陽性の結果が出た場合、この女性ががんである確率は何か？

(a) 0.0936
(b) 0.088
(c) 0.48
(d) 0.52

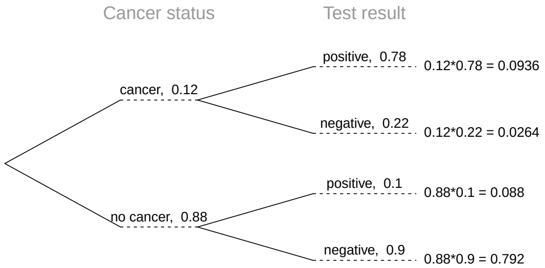


$$P(C|+) = \frac{P(C \text{ かつ } +)}{P(+)} = \frac{0.0936}{0.0936 + 0.088} = 0.52$$

練習問題

この 2 回目のマンモグラフィでも陽性の結果が出た場合、この女性ががんである確率は何か？

(a) 0.0936
(b) 0.088
(c) 0.48
(d) *0.52*



$$P(C|+) = \frac{P(C \text{ かつ } +)}{P(+)} = \frac{0.0936}{0.0936 + 0.088} = 0.52$$

ベイズの定理

- これまで見てきた条件付き確率の公式は、ベイズの定理の特殊な場合であり、事象が 2 つ以上の結果を持つ場合にも適用できる。
- ベイズの定理：

$$P(\text{変数 1 の結果 } A_1 \mid \text{変数 2 の結果 } B)$$
$$= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)}$$

ここで $A_2, \cdots, A_k$ は変数 1 のその他のすべての可能な結果を表す。

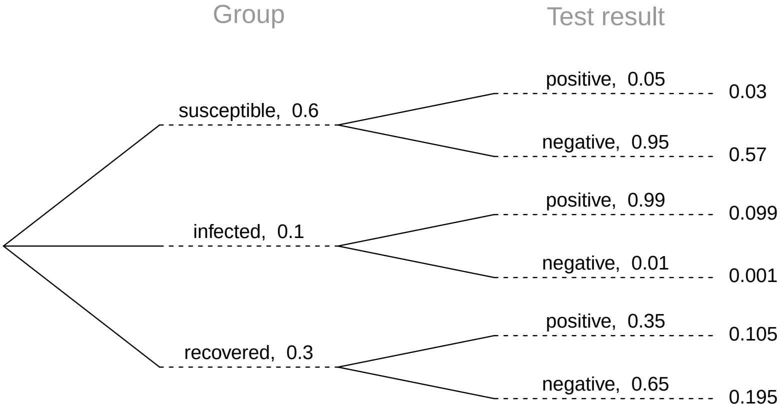
応用活動：確率の逆転

疾病の拡散を表す一般的な疫学モデルとして SIR モデルがある。このモデルでは、人口が感受性（Susceptible）、感染（Infected）、回復（Recovered）の 3 つのグループに分けられる。これは、1 回の感染が通常その後の感染に対する免疫を与える水痘のような疾患に対して合理的なモデルである。これらの疾患は検出が難しいこともある。

人口の 60%が感受性、10%が感染、30%が回復状態にある疫病の最中の人口を想像する。この疾患の唯一の検査は、感受性者に対して 95%、感染者に対して 99%、回復者に対して 65%の精度を持つ。（注：ここで精度とは、感受性者と回復者に対して陰性、感染者に対して陽性の結果を返すことを意味する）。

上記の情報を反映した確率樹を描きなさい。個人が陽性と検査された場合、実際に感染している確率は何か？

応用活動：確率の逆転（続き）



$$P(\text{感染}|+) = \frac{P(\text{感染 かつ }+)}{P(+)} = \frac{0.099}{0.03 + 0.099 + 0.105} \approx 0.423$$

復元抽出

復元抽出では、引いたものを元に戻す。

- 赤チップ 5 個、青チップ 3 個、オレンジチップ 2 個が入った袋があるとする。最初に引いたチップが青である確率は何か？

5 ●, 3 ●, 2 ●

$$Prob(1\text{番目のチップが青}) = \frac{3}{5+3+2} = \frac{3}{10} = 0.3$$

- 1 回目の引きで青チップを引いたとする。復元抽出で引く場合、2 回目の引きで青チップを引く確率は何か？

1回目: 5 ●, 3 ●, 2 ●
2回目: 5 ●, 3 ●, 2 ●

$$Prob(2\text{番目のチップが青}|1\text{番目のチップが青}) = \frac{3}{10} = 0.3$$

復元抽出（続き）

- 1 回目の引きでオレンジチップを引いたとする。復元抽出で引く場合、2 回目の引きで青チップを引く確率は何か？

1回目: 5 ●, 3 ●, 2 ●

2回目: 5 ●, 3 ●, 2 ●

$$Prob(2\text{番目のチップが青}|1\text{番目のチップがオレンジ}) = \frac{3}{10} = 0.3$$

- 復元抽出で、2 回続けて青チップを引く確率は何か？

1回目: 5 ●, 3 ●, 2 ●

2回目: 5 ●, 3 ●, 2 ●

$$Prob(1\text{番目のチップが青}) \cdot Prob(2\text{番目のチップが青}|1\text{番目のチップが青})$$
$$= 0.3 \times 0.3$$
$$= 0.3^2 = 0.09$$

復元抽出（続き）

- 復元抽出では、2 番目のチップが青である確率は、1 番目のチップの色に依存しない。なぜなら、1 回目に引いたものを袋に戻すからである。

$$Prob(\text{青}|\text{青}) = Prob(\text{青}|\text{オレンジ})$$

- さらに、この確率は 1 番目のチップが青である確率と等しい。なぜなら、復元抽出では袋の構成が変わらないからである。

$$Prob(\text{青}|\text{青}) = Prob(\text{青})$$

- 復元抽出では、引き操作は独立である。

非復元抽出

非復元抽出では、引いたものを元に戻さない。

- 1 回目の引きで青チップを引いたとする。非復元抽出で引く場合、2 回目の引きで青チップを引く確率は何か？

1回目: 5 ●, 3 ●, 2 ●
2回目: 5 ●, 2 ●, 2 ●

$$Prob(2\text{番目のチップが青}|1\text{番目のチップが青}) = \frac{2}{9} = 0.22$$

- 非復元抽出で、2 回続けて青チップを引く確率は何か？

1回目: 5 ●, 3 ●, 2 ●
2回目: 5 ●, 2 ●, 2 ●

$$Prob(1\text{番目のチップが青}) \cdot Prob(2\text{番目のチップが青}|1\text{番目のチップが青})$$
$$= 0.3 \times 0.22$$
$$= 0.066$$

非復元抽出（続き）

- 非復元抽出では、1 番目が青であった場合の 2 番目のチップが青である確率は、1 番目のチップが青である確率と等しくない。なぜなら、袋の構成が 1 回目の引きの結果によって変わるからである。

$$Prob(\text{青}|\text{青}) \neq Prob(\text{青})$$

- 非復元抽出では、引き操作は独立でない。
- これは、標本サイズが小さいときに特に注意が必要である。たとえば、（巨大な）袋に 10,000 個のチップが入っている場合、1 つのチップを引き出しても 2 回目の引きの確率にはそれほど大きな影響を与えない。

練習問題

ほとんどのカードゲームでは、カードは非復元で配られる。エースを引いてから 3 を引く確率は何か？ 最も近い答えを選びなさい。

(a) 0.0045
(b) 0.0059
(c) 0.0060
(d) 0.1553

$$P(\text{エース次に }3) = \frac{4}{52} \times \frac{4}{51} \approx 0.0060$$

練習問題

ほとんどのカードゲームでは、カードは非復元で配られる。エースを引いてから 3 を引く確率は何か？ 最も近い答えを選びなさい。

(a) 0.0045
(b) 0.0059
(c) *0.0060*
(d) 0.1553

$$P(\text{エース次に }3) = \frac{4}{52} \times \frac{4}{51} \approx 0.0060$$

確率変数

- 確率変数とは、確率的事象の結果に依存する数値量のことである
  - 確率変数を表すために X のような大文字を使う
  - 確率変数の値は小文字、ここでは x で表す
  - 例：$P(X = x)$
- 確率変数には 2 種類ある：
  - 離散確率変数はしばしば整数値のみをとる
    - 例：履修単位数、今学期と前学期の履修単位数の差
  - 連続確率変数は実数（小数）値をとる
    - 例：今学期の教科書代、今学期と前学期の教科書代の差

期待値

- 確率変数の平均的な結果にしばしば関心がある。
- これを期待値（平均）と呼び、可能な結果の加重平均である

$$\mu = E(X) = \sum_{i=1}^{k} x_i \; P(X = x_i)$$

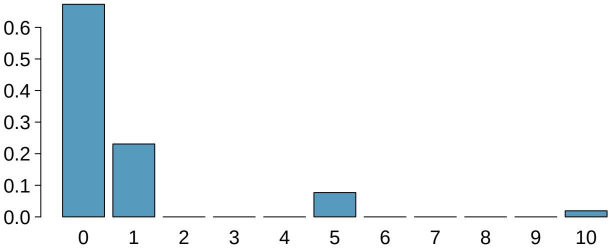
離散確率変数の期待値

カードゲームで、ハートを引いたら 1 ドル、エース（ハートのエースを含む）を引いたら 5 ドル、スペードのキングを引いたら 10 ドル、それ以外のカードを引いたら何も得られない。勝利金の確率モデルを書き、期待される勝利金を計算しなさい。

事象	X	$P(X)$	$X\ P(X)$
ハート（エース以外）	1	$\frac{12}{52}$	$\frac{12}{52}$
エース	5	$\frac{4}{52}$	$\frac{20}{52}$
スペードのキング	10	$\frac{1}{52}$	$\frac{10}{52}$
その他	0	$\frac{35}{52}$	0
合計			$E(X) = \frac{42}{52} \approx 0.81$

離散確率変数の期待値（続き）

以下はこのゲームの勝利金の確率分布を視覚的に表したものである：



変動性

確率変数の値の変動性にもしばしば関心がある。

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

離散確率変数の変動性

先のカードゲームの例で、ゲームごとに勝利金がどのくらい変動すると予想されるか？

X	$P(X)$	$X\ P(X)$	$(X - E(X))^2$	$P(X)\ (X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246} = 1.85$

線形結合

- 確率変数 X と Y の線形結合は次のように表される

$$aX + bY$$

  ここで a と b はある固定された数である。
- 確率変数の線形結合の平均値は次のように計算される

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

線形結合の期待値の計算

統計学の宿題問題 1 つには平均 10 分、化学の宿題問題 1 つには平均 15 分かかる。今週は統計学の宿題 5 問と化学の宿題 4 問が出された。今週の統計学と化学の宿題に費やすと予想される合計時間は何か？

$$
\begin{aligned}
&E(S+S+S+S+S+C+C+C+C) \\
&= 5 \times E(S) + 4 \times E(C) \\
&= 5 \times 10 + 4 \times 15 \\
&= 50 + 60 \\
&= 110 \text{ 分}
\end{aligned}
$$

線形結合

- 2つの独立な確率変数の線形結合の変動性は次のように計算される

$$V(aX+bY)=a^2\times V(X)+b^2\times V(Y)$$

- 線形結合の標準偏差は分散の平方根である。

注：確率変数が独立でない場合、分散の計算はやや複雑になり、本講義の範囲を超える。

線形結合の分散の計算

統計学の宿題問題1つにかかる時間の標準偏差は1.5分、化学の宿題問題1つには2分である。統計学の宿題5問と化学の宿題4問が出された今週に統計学と物理学の宿題に費やすと予想される時間の標準偏差は何か？ 各問題にかかる時間は互いに独立であると仮定する。

$$
\begin{aligned}
&V(S+S+S+S+S+C+C+C+C) \\
&= V(S)+V(S)+V(S)+V(S)+V(S) \\
&+ V(C)+V(C)+V(C)+V(C) \\
&= 5 \times V(S) + 4 \times V(C) \\
&= 5 \times 1.5^2 + 4 \times 2^2 \\
&= 27.25
\end{aligned}
$$

練習問題

カジノゲームのプレイに 5 ドルかかる。最初に引いたカードが赤であれば、2 枚目のカードを（非復元で）引ける。2 枚目がクラブのエースであれば 500 ドルを獲得する。そうでなければ、何も得られず（5 ドルを失う）。このゲームをプレイしたときの期待利益/損失は何か？ 覚えておこう：利益/損失 = 賞金 - コスト。

(a) 5 セントの利益
(b) 10 セントの損失
(c) 25 セントの損失
(d) 30 セントの損失

事象	賞金	利益: X	$P(X)$	$X \times P(X)$
赤, A♣	*500*	*500 - 5 = 495*	$\frac{26}{52} \times \frac{1}{51} = 0.0098$	$495 \times 0.0098 = 4.851$
その他	*0*	*0 - 5 = -5*	$1 - 0.0098 = 0.9902$	$-5 \times 0.9902 = -4.951$
				$E(X) = -0.1$

練習問題

カジノゲームのプレイに 5 ドルかかる。最初に引いたカードが赤であれば、2 枚目のカードを（非復元で）引ける。2 枚目がクラブのエースであれば 500 ドルを獲得する。そうでなければ、何も得られず（5 ドルを失う）。このゲームをプレイしたときの期待利益/損失は何か？ 覚えておこう：利益/損失 = 賞金 - コスト。

(a) 5 セントの利益
(b) *10 セントの損失*
(c) 25 セントの損失
(d) 30 セントの損失

事象	賞金	利益: X	$P(X)$	$X \times P(X)$
赤, A♣	*500*	*500 - 5 = 495*	$\frac{26}{52} \times \frac{1}{51} = 0.0098$	$495 \times 0.0098 = 4.851$
その他	*0*	*0 - 5 = -5*	$1 - 0.0098 = 0.9902$	$-5 \times 0.9902 = -4.951$
				$E(X) = -0.1$

公正なゲーム

公正なゲームとは、コストが期待される賞金額と同じ、すなわち期待利益が 0 のゲームである。

ラスベガスのカジノゲームは、期待される賞金額より多くかかると思うか、少なくかかると思うか？

もしそれらのゲームが期待される賞金額より少なければ、カジノは平均的に損をしていることになり、以下のようなものに費用を払えなくなる：

Image by Moyan_Brenn on Flickr http://www.flickr.com/photos/aigle_dore/5951714693.

確率変数の簡略化

確率変数は通常の代数変数とは異なる振る舞いをする：

$$X + X \neq 2X$$

$$\begin{aligned} E(X+X) &= E(X)+E(X) \\ &= 2E(X) \end{aligned}$$

$$\begin{aligned} Var(X+X) &= Var(X)+Var(X) \quad \text{（独立を仮定）} \\ &= 2\ Var(X) \end{aligned}$$

$$E(2X) = 2E(X)$$

$$\begin{aligned} Var(2X) &= 2^2\ Var(X) \\ &= 4\ Var(X) \end{aligned}$$

$E(X+X) = E(2X)$ だが、$Var(X+X) \neq Var(2X)$ である。

足し算か掛け算か？

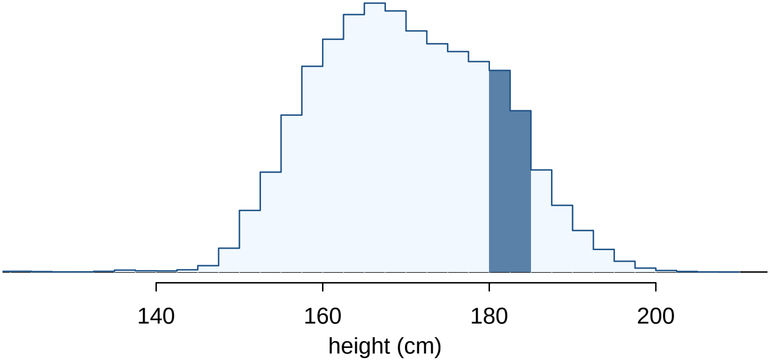
ある会社のフリートに 5 台のリンカーン・タウンカーがある。過去のデータによると、各車の年間メンテナンスコストは平均 2,154 ドル、標準偏差は 132 ドルである。このフリートの年間総メンテナンスコストの平均と標準偏差は何か？

1 台の車が年間メンテナンスコストの 5 倍 $(5X)$ ではなく、それぞれ同じ年間メンテナンスコストを持つ 5 台の車 $(X_1 + X_2 + X_3 + X_4 + X_5)$ であることに注意する。

$$
\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var(X_5) \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120 \\
SD(X_1 + X_2 + X_3 + X_4 + X_5) &= \sqrt{87,120} = 295.16
\end{aligned}
$$

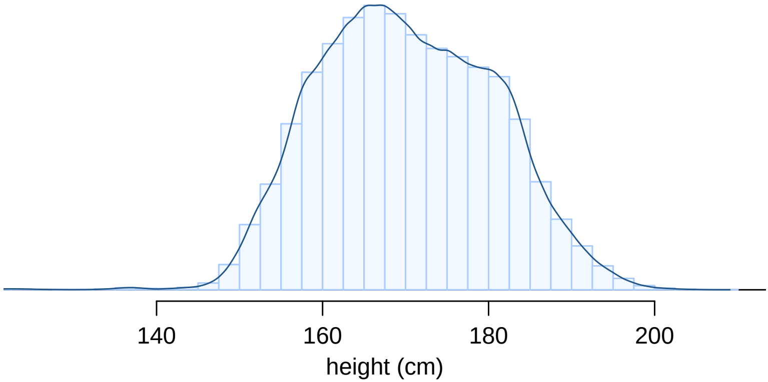
連続分布

- 以下は米国成人の身長の分布のヒストグラムである。
- 陰影のついたビンに含まれるデータの割合が、無作為に選ばれた米国成人の身長が 180cm から 185cm（約 5’11” から 6’1”）である確率を示す。



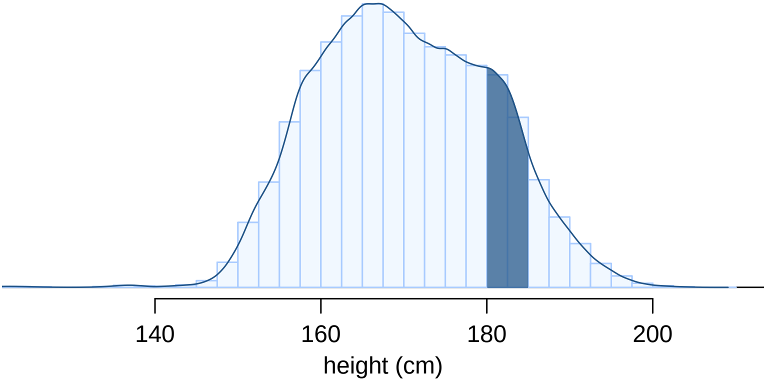
ヒストグラムから連続分布へ

身長は連続数値変数であるため、その確率密度関数は滑らかな曲線となる。



連続分布からの確率

したがって、無作為に選ばれた米国成人の身長が 180cm から 185cm である確率は、曲線の下の陰影部分の面積としても推定できる。



定義によれば…

連続確率は「曲線の下の面積」として推定されるため、ある人の身長がちょうど 180cm（または任意の特定の値）である確率は 0 と定義される。