1章　序論

東京大学

2014

Outline

Outline

Outline

Outline

Outline

統計学とは

一言で言えば …

データを収集し、それを分析する学問

データから何を分析するのか？

母集団 (population): データの抽出元、データを生成する構造自体

統計学の目的：

母集団そのものの性質を調べること

例: 日本の内閣支持率を考える。「母集団」は？

- 母集団＝有権者全員。
- データ（「標本」）＝有権者の一部に対する聞き取り調査の結果。

記述統計と推測統計

記述統計 :
データを分かりやすく記述する。

推測統計:
データから母集団の特性を推測する。

Remark

全数調査が可能であれば、データは母集団と一致し、その特性を直接知ることができる。しかし、多くの場合、全数調査は時間とコストの両面から困難であり、推測統計が必要となる。

例: テレビの視聴率

データ

例：身長のデータ

$$(X_1, X_2, X_3, X_4, X_5) = (172.9, 180.3, 142.1, 120.2, 172.3)$$

変数の種類：

連続変数：
連続的な値をとりうる変数：身長、体重 ···

離散変数:
離散的な値だけを取りうる変数：ある試合で勝ち負けを記録したデータ（勝ったら 1、負けたら 0）、内閣支持率調査のデータ（支持すると答えたら 1、支持しないと答えたら 0）

データの種類：

時系列データ (time-series data)：時間の経過とともに観測されるデータ。例えば、1974 ~ 2011 年の東京都の県内総生産の記録。

横断面データ (cross-section data):
ある 1 時点において複数の対象を記録したデータ。例えば、1975 年だけの 47 都道府県の県内総生産。

パネルデータ (panel data):
横断面データが複数年にまたがって利用できる場合。例えば、
1974 ~ 2011 年の 47 都道府県の県内総生産の記録。

データの収集

データは、できるだけ**偏り (バイアス)** なく、**母集団を代表するデータ**を取り出す必要がある。

- バイアスのあるデータの分析からは、バイアスのある結果が出てしまう。
- 例えば、大学生の意識調査をする時に、友人にだけ聞取り調査をすると、母集団である大学生を代表していない可能性がある。なぜか？

「類を友を呼ぶ」

バイアスの発生を回避する方法：**無作為抽出 (random sampling)**=母集団を構成するどの個体もデータとして選ばれる確率が同じになる抽出法。

- 現実には誤ったデータ抽出が頻繁に行われている。
- 例えば、1936 年の米国大統領選において、『リテラシー・ダイジェスト』誌が勝利者を予想するため、電話や自動車の保有者などから選ばれた約 237 万人に聞取り調査を行った結果、共和党候補ランドン氏が圧倒的な優勢となった。しかし、実際の選挙では民主党候補ルーズベルト氏の勝利となった。なぜ調査結果は誤ったか？

Remark

社会調査では、無作為抽出のほかにも、*1.* 面接員の質問の仕方、*2.* 質問の設定の仕方、*3.* 回答者が嘘をつく可能性等にも、注意を払う必要がある。(教科書 *1.3.2* 節参照)

確率の計算

直観は頼りになるか？

例：HIV 検査の偽陽性問題

HIV 検査において、感染者については 100% の確率で陽性反応が出るが、非感染者でも検査に反応する抗体を持っている可能性があり、1% の確率で陽性反応を示すとする。全人口の 0.1% だけが感染者であると仮定した場合、陽性の検査結果が出た時、その人が HIV に感染している確率はどのくらいだろうか？

確率の計算

直観は頼りになるか？

例：HIV 検査の偽陽性問題
　HIV 検査において、感染者については 100% の確率で陽性反応が出るが、非感染者でも検査に反応する抗体を持っている可能性があり、1% の確率で陽性反応を示すとする。全人口の 0.1% だけが感染者であると仮定した場合、陽性の検査結果が出た時、その人が HIV に感染している確率はどのくらいだろうか？

9%

仮説検定 (hypothesis testing)

仮説がデータと整合的かどうかを検証する方法。

- **「帰無仮説」 (null hypothesis)**: 検証したい仮説。
- **「対立仮説」 (alternative hypothesis)**: 帰無仮説が誤っていたときの受け皿としての仮説。
- 仮説検定は、これらの仮説を設定した上で、どちらの仮説がデータと整合的であるかを判断する。

例：太郎君が、「気になる女性がある社内の別の男性と恋愛中であるかどうか」をその二人が有給休暇を取っているタイミングの「データ」をもとに「仮説検定」したい。

- 帰無仮説：「二人の間に恋愛関係がない」
- 対立仮説：「二人の間に恋愛関係がある」
- 二人の有給休暇を調べた結果、タイミングが完全に一致していた。
- 太郎君は、これを偶然と考えるのには無理があると判断し、「二人の間に恋愛関係がある」、つまり「対立仮説」が正しいと判断した。

2 章 データの記述

東京大学

2014

Outline

Outline

図表の作成

例：ある工場における従業員の欠勤日数のデータ（一年間）

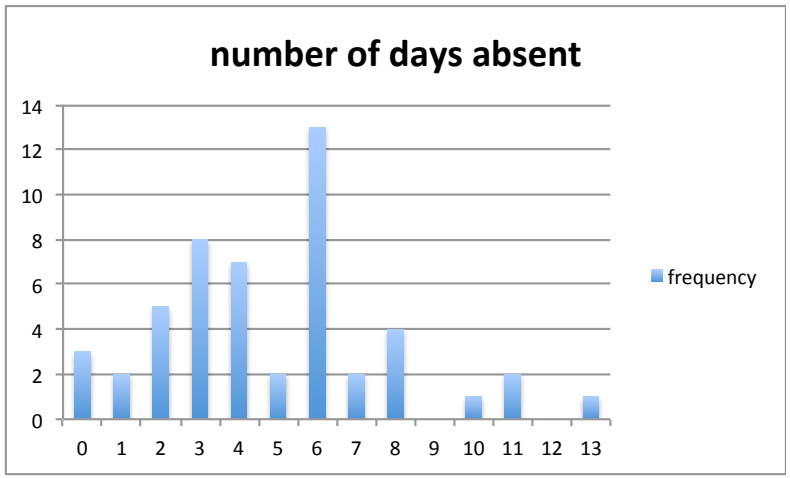
$$
\begin{aligned}
(X_1, \ldots, X_{50}) \quad = \quad & 6,4,4,6,0,\ 6,11,5,10,8,\ 4,8,4,7,7,\ 3,2,3,6,2 \\
& 4,3,6,1,3,\ 2,4,6,6,6,\ 6,8,3,3,6,\ 2,3,2,4,0 \\
& 8,3,6,0,1,\ 6,5,13,11,6
\end{aligned}
$$

- 度数分布表 (frequency table)
- ヒストグラム (histogram)

度数分布表：データの大きさによっていくつかの組（階級）に分けて、**度数** (各組に入る観測値の数）をまとめた表

- 階級値：観測値が属する各組の中心値
- 相対度数：総度数に対する各度数の割合
- 累積相対度数：相対度数を下から順に加えてその累積値を求めたもの

ヒストグラム：各階級の度数を長方形によりグラフ表示したもの。



- 特にデータが**連続変数**（身長、所得など）の場合、分析者によって階級数を決めなくてはならない。
- 階級の決め方は目的によって異なる。

例：CPS (Current Population Survey, 2008) data in U.S.

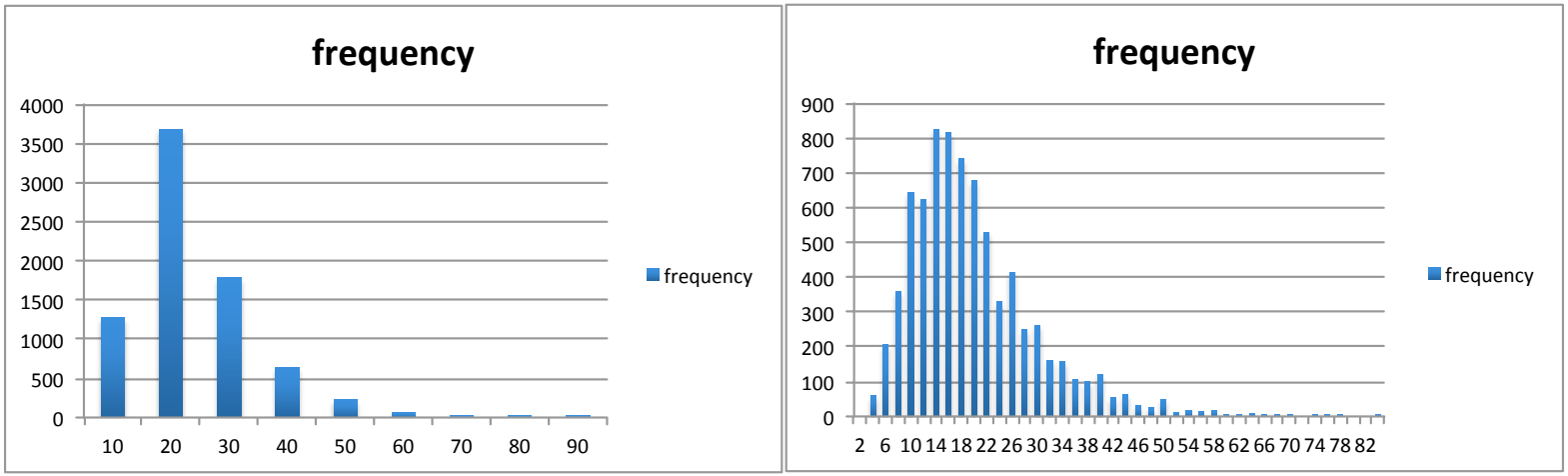


Figure : Hourly earnings in the United States of Working Graduates, Ages 28-34 in 2008 Dollars

代表的分布の形状

- **釣鐘状の分布**：左右対称な釣鐘状の分布。(適切に設計された試験の成績、身長、株価の変化率など)
- **右に歪んだ分布**：右裾が長く、左裾が短い分布。(所得、資産、結婚年齢、体重など)
- **左に歪んだ分布**：左裾が長く、右裾が短い分布。(簡単な試験の成績、人間の寿命など)

中心を表す標本特性値

1. 平均 (mean)

以下、データを $(x_1, x_2, \ldots, x_n)$ とする。平均は以下で定義される。

平均

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 平均には、**外れ値**の影響を受けやすいという欠点がある。
- 例えば、$\{1, 2, 3, 4, 5\}$ というデータが与えられた場合、平均は $(1+2+3+4+5)/5 = [\quad]$。
- ここで、何らかの原因で 5 の代わりに 90 という外れ値がデータに含まれているとする。このとき、新しいデータ $\{1, 2, 3, 4, 90\}$ の平均は $(1+2+3+4+90)/5 = [\qquad]$ となり、外れ値の影響を大きく受けることになる。

2. 中央値 (median)

まず、データを小さい順（昇順）に並べ直す（$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$）。
中央値（median）は以下で定義される。

中央値

n が奇数なら：$x_{(n+1)/2}$　　n が偶数なら：$\frac{x_{(n/2)}+x_{(n/2+1)}}{2}$

- 中央値は、**外れ値**の影響を受けない。
- $\{1, 2, 3, 4, 5\}$ というデータの中央値は [　]。
- $\{1, 2, 3, 4, 90\}$ の中央値は [　]。

3. 最頻値 (mode)

最頻値（mode）は、

最頻値

「最大の頻度を持つ測定値」

として定義される。

- 度数分布表が与えられている時には（階級幅が同じという条件の下で）最大度数を与える階級の階級値を最頻値とする。
- 最頻値も**外れ値**の影響を受けない。

中心を表す3つの特性値の使い方

平均、中央値、最頻値の3つの値を相互比較することによって有用な情報が得られる。

- 分布が釣鐘状なら、[　　=　　=　　]
- 分布が右に歪んでいるなら、[　　<　　<　　]
- 分布が左に歪んでいるなら、[　　<　　<　　]

従って、特性値を相互比較することによって、分布の形状をおおまかに推測できる。

ばらつきを表す標本特性値

例：試験の成績

グループ A：40 点、42 点、58 点、60 点
グループ B：20 点、35 点、45 点、100 点

- 両グループとも、平均は 50 点。
- グループ A は点数のばらつきが小さく、B はばらつきが大きい。
- A は平均を中心に狭い範囲でばらつき、B は平均を中心に広範囲に散らばっている。

$\longrightarrow$ ばらつきの指標を作るために、平均からの乖離を考える。

- ID 番号 i の人の点数 x_i の平均 $\bar{x}$ からの乖離 $x_i - \bar{x}$ を**偏差 (deviation)** と呼ぶ。
- しかし、全て偏差を足し合わせるだけでは、全体のばらつきを測ることはできないことに注意する。($\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.)

$\longrightarrow$ 「偏差の二乗を足し合わせたもので、全体のばらつきを測る」という考え方による指標。

標本分散 (sample variance)

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

- 除数をサンプルサイズ n ではなく、$n-1$ としている理由は後ほど説明。
- 標本分散は偏差の二乗をもとに求めた値であるため、桁数がデータの桁数から変わってしまう。
- データのばらつきの直観をつかむために、標本分散の平方根をとったものを**標本標準偏差 (sample standard deviation)** と呼ぶ。

標本標準偏差 (sample standard deviation)

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

先ほどの点数の例では、グループ A は、

$$s_x^2 = [\qquad\qquad]$$
$$s_x = [\qquad\qquad]$$

グループ B は、

$$s_x^2 = [\qquad\qquad]$$
$$s_x = [\qquad\qquad]$$

データの線形変換

例：入試の得点調整

世界史：平均 70 点、標準偏差 15 点
日本史：平均 50 点、標準偏差 10 点

- このままでは世界史の平均が高いため、日本史で受験した学生に不利になる可能性あり。
- そこで、日本史の得点の平均と標準偏差を世界史のそれに合わせるという得点調整を考える。
- 具体的には、日本史の点数が x_i のとき、$y_i = a + bx_i$ と**線形変換**し、得点調整後の日本史の得点 y_i の平均 $(\bar{y})$ が 70 点、標準偏差 (s_y) が 15 点となるように a,b を選ぶ。

ここで、次の関係が成り立つことに注意する。

$$\bar{y} = a + b\bar{x}, \quad s_y^2 = b^2 s_x^2, \quad s_y = |b| s_x$$

今、$\bar{x} = 50$、$s_x = 10$ なので、$\bar{y} = 70$, $s_y = 15$ となるためには、連立方程式

$$70 = a + 50b, \quad 15 = 10b$$

を解いて、$a =$ [　　], $b =$ [　　] とすればよい。

線形変換のイメージ

- $y_i = a + x_i,\ a > 0$ の場合：分布の [　　] は変わらないが、[　　] だけ [　　] にシフトした分布になる。
- $y_i = bx_i,\ b > 1$ の場合：分布の [　　] は変わらないが、[　　] が [　　] なった分布になる。
- $y_i = bx_i,\ 0 < b < 1$ の場合：分布の [　　] は変わらないが、[　　] が [　　] なった分布になる。

- $y_i = bx_i,\ b = -1$ の場合：x の分布の [　　] した分布となる。
- $y_i = bx_i,\ -1 < b < 0$ の場合：x の分布を [　　] し、[　　] が [　　] なった分布になる。
- $y_i = bx_i,\ b < -1$ の場合：x の分布を [　　] し、[　　] が [　　] なった分布になる。

範囲と割合の関係

経験的に（厳密にはデータが「正規分布」と呼ばれる分布に従う場合）、範囲と割合の間には次のような関係が成り立つことが知られている。

範囲	割合
$\bar{x} \pm s_x$	約 68%
$\bar{x} \pm 2s_x$	約 95%
$\bar{x} \pm 3s_x$	約 99 ～ 100%

例：偏差値

$$i \text{ さんの偏差値} = 50 + 10\left(\frac{x_i - \bar{x}}{s_x}\right)$$

偏差値 40～60 の間に全体の約 68%、偏差値 30～70 の間に全体の約 95%、偏差値 20～80 の間に全体の約 99～100% が入る。

3 章 相関

東京大学

2014

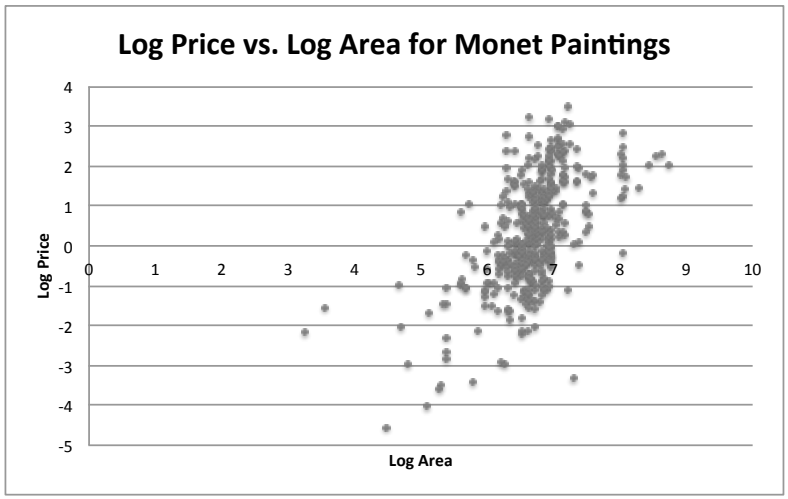
Outline

Outline

散布図

散布図 (scatter plot): 2 変数からなるデータ (x, y) を平面上の点としてプロットしたもの。

例：モネ (Monet) の絵画オークションのデータ



散布図を作ることによって、2 変数の大まかな関係が分かる。

散布図から分かる 3 つの関係：

- **正の相関がある場合**：x が高いと y も高い場合（教育と賃金の関係など）
- **負の相関がある場合**：x が高いと y が低い場合（喫煙量と寿命の関係、駅からの距離と地価の関係など）
- **相関がない場合**：x と y の動きに関連性が見られない場合

例：為替レートの変化率

標本共分散

標本共分散 (sample covariance) は、2 変数の共変動を表す指標であり、標本共分散がプラスなら**正の相関**、マイナスなら**負の相関**、0 なら**無相関**、0 に近いなら**相関が弱い**とされる。

標本共分散 (sample covariance)

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

例：太郎君、次郎君、三郎君の身長 x と体重 y が次のとおりだったとする。

$$(180\text{cm},\ 80\text{kg}),\quad (170\text{cm},\ 70\text{kg}),\quad (160\text{cm},\ 60\text{kg})$$

この時、$\bar{x} = [\qquad\qquad]$、$\bar{y} = [\qquad\qquad]$ なので、

$s_{xy} = [\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad]$

標本共分散の欠点：スケール（尺度）が変わると、その値も変わってしまう。

- たとえば、cm 表示から m 表示に変更すると、標本共分散は [　　] 倍されてしまう。

$\longrightarrow$ この問題を解決する指標が**標本相関係数 (sample correlation coefficient)**

標本相関係数 (sample correlation coefficient)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- 標本相関係数は、標本共分散と符号は同じ。
- 標本相関係数は、スケール変更に依存しない。
- 標本相関係数は、-1 から 1 の間の値を取る。
- 標本相関係数が 1 または-1 となるのは、変数 x と y に完全な**線形関係** ($y = a + bx$) が成立しているとき。

例：先ほどの 3 兄弟の身長と体重の標本相関係数は

$r_{xy} = [$ $]$

標本相関係数の注意点

1. 相関と因果関係は異なる概念である。
   - **相関**は両変数の動きに関連性があるかを示す概念であるのに対し、**因果関係**は原因と結果という関係を意味する概念である。
   - x と y に相関があっても、それは x から y への因果なのか、y から x への因果なのか、あるいは両方の因果が混ざっているのかは分からない。

(両方の因果が混ざっている場合の例)：選挙資金と選挙結果の間には正の相関があるが、「選挙資金が当選を決める重要な要因である」とはいえない。

(相関があっても因果関係を全く意味しない例「**見せかけの相関**」)：溺死者数とアイスクリーム消費量の間には正の相関があるが、それは「夏の暑さ」によって生まれたものであり、両者に因果関係はない。

2. 標本相関係数でとらえることができるのは**線形関係**だけである。

例: 電力需要と気温の関係は U 字型となる。そのため、相関係数では変数間の相互関係を上手くとらえることはできない。

4 章 確率

東京大学

2014

Outline

Outline

推測統計は確率論に基づいて行われる。また、確率論の基礎は集合論である。ここでは、まず集合論の基本概念を概観する。

Definition

- 結果が偶然に支配される実験を**試行 (trial)** という。
- 試行により生じる実現可能な全ての結果を集めたものを**標本空間 (sample space)** といい、Ω で表す。
- 標本空間の個々の異なった結果を**標本点 (sample point)** といい、ω で表す。
- 標本空間の部分集合を**事象 (event)** という。

Example 1: コイン投げの試行の場合、 標本空間は表と裏, すなわち $\Omega = \{H, T\}$.

Example 2: 「一組の夫婦が三人の子供を産む」という試行の場合、一つの標本点は例えばBGB、すなわち「男の子、次に女の子、最後に男の子」となる。標本空間は、

$\Omega = \{$ $\}$. (2^3通

Example 1, continued: コイン投げにおいて、可能な全ての事象 (all possible events) は

$$[H, T, \Omega, \phi] .$$

である。ここで、ϕ は**空事象 (null event)** という。

二つの事象 A と B が与えられた場合、次のように集合演算を定義する。

Definition

和事象 (Union): A と B の少なくとも一方が起こる事象

$$A \cup B = \{x : x \in A \quad or \quad x \in B\}.$$

積事象 (Intersection): A と B が同時に起こる事象

$$A \cap B = \{x : x \in A \quad and \quad x \in B\}.$$

余事象 (Complement): A の余事象は、「A が起こらない事象」

$$\bar{A} = \{x : x \notin A\}.$$

Example 3: トランプからカードを取り出す試行において、カードのマーク (club(C), diamonds(D), hearts(H), or spades(S)) に注目する。標本空間は

$$\Omega = \{ \qquad \},$$

であり、事象には、例えば

$$A = \{C, H\}, \qquad B = \{C, D, S\}.$$

が含まれる。このとき、

$$A \cup B = \{ \qquad \}, \ A \cap B = \{ \qquad \}, \ \bar{A} = \{ \qquad \}.$$

「確率」を 以下のように定義する。

Definition

ある事象 A が起こる確率とは、観察回数 n と事象 A が起こった回数 $n(A)$ との相対度数 $n(A)/n$ の極限値である。

$$Pr(A) = \lim_{n \to \infty} \frac{n(A)}{n}.$$

例えば、サイコロの各目が出る確率は、サイコロを何度も投げた後、その目が観測される相対度数の極限値（相対度数が近づいていく値）として定義される。歪みのないサイコロであれば、各目の確率は 1/6 となる。

The following is more rigorous definition of "probability", which is optional.

Definition

Definition(*optional*): *A collection of subsets of* Ω *is called a* **sigma algebra** *(or Borel field), denoted by* $\mathcal{B}$*, if it satisfied the following three properties:*

a. $\phi \in \mathcal{B}$*, where* ϕ *is the empty set.*

b.If $A \in \mathcal{B}$*, then* $\bar{A} \in \mathcal{B}$*.*

c.If $A_1, A_2, \ldots, \in \mathcal{B}$*, then* $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$*.*

Definition

Definition(*optional*): *Given a sample space* Ω *and associated sigma algebra* $\mathcal{B}$*, a* **probability function** *is a function* P *with domain* $\mathcal{B}$ *that satisfies*

1. $P(A) \geq 0$ *for all* $A \in B$*.*
2. $P(\Omega) = 1$*.*
3. *If* $A_1, A_2, \ldots, \in \mathcal{B}$ *are mutually exclusive (i.e., do not overlap), then* $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$*.*

Example 2, continued: 一組の夫婦が 3 人の子供を産む例において、すべての出産において男の子が生まれる確率と女の子が生まれる確率がそれぞれ 50% であるとする。この時、「男の子 → 女の子 → 男の子」となる「確率」は？

$$Pr(BGB) = [\qquad\qquad\qquad\qquad\qquad\qquad].$$

夫婦が次のような事象を望んでいるとする。

$$E = \text{少なくとも 2 人は女の子}$$

この時、

$$Pr(E) = [\qquad\qquad\qquad\qquad\qquad\qquad]. \ (1/2)$$

次のような事象を考える。

$$
\begin{aligned}
F &= \text{少なくとも 1 人は女の子} \\
G &= \text{女の子は 2 人未満} \\
H &= \text{全ての子の性別が同じ} \\
K &= \text{男の子が 2 人未満} \\
I &= \text{女の子なし}
\end{aligned}
$$

$Pr(I \cup K) = [$ $]. (5/8)$

$Pr(G \cup H) = [$ $]. (5/8)$

$Pr(F) = [$ $]. (7/8)$

便利な公式：

$$
\begin{aligned}
Pr(A \cup B) &= Pr(A) + Pr(B) - Pr(A \cap B). \\
Pr(A) &= 1 - Pr(\bar{A}).
\end{aligned}
$$

事象 A と B が互いに共通点を持たない (片方の事象が起こったら、もう片方は絶対に起こらないような) 時、二つの事象は互いに**排反 (disjoint or mutually exclusive)** であるという。

例えば、事象 I と事象 K は排反である。

便利な公式：

事象 A と B が排反なら、

$$Pr(A \cup B) = Pr(A) + Pr(B)$$

条件付確率 (Conditional Probability)

Example 2, continued：3 人の子供を産む例において、事象 G (女の子が二人未満) が起こったとする。この時、事象 H (全ての子の性別が同じ) となる確率は？

この確率を「事象 G が起こったという条件のもとで、事象 H が起こる**条件付確率 (conditional probability)**」といい、次のように定義される。

Definition

事象 G が起こったという条件のもとで、事象 H が起こる条件付確率 *(conditional probability)*

$$Pr(H|G) = \frac{Pr(H \cap G)}{Pr(G)}.$$

(a) 男の子が生まれる確率が 50% の時、$Pr(H|G)$ は？

$$Pr(H|G) = [\qquad\qquad\qquad\qquad\qquad\qquad] \ (1/4)$$

(b) 男の子が生まれる確率が 52% の時、$Pr(H|G)$ は？

$$Pr(H|G) = [\qquad\qquad\qquad\qquad\qquad\qquad] \ (0.26)$$

（例）お隣さんの飼っている犬が 2 匹の子犬を産みました。お隣さんは「少なくとも 1 匹はオスでしたよ」と教えてくれました。この時、残りの犬もオスである確率は何 % でしょうか？（オスが産まれる確率は 50% であるとする。）

（答）2 匹ともオスである場合を事象 A、少なくとも 1 匹がオスである場合を事象 B とする。

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(MM)}{Pr(MM, MF, FM)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

条件付確率の定義から、次の公式が成り立つ。

Theorem

$$Pr(G \cap H) = Pr(G)Pr(H|G)$$

Example 4: Suppose that 3 defective light bulbs inadvertently got mixed with 6 good ones. If 2 bulbs are chosen at random for a ceiling lamp, what is the probability that they both are good?

Answer: Let us denote

$$\begin{aligned} G_1 &= \text{first bulb is good} \\ G_2 &= \text{second bulb is good} \end{aligned}$$

Thus,

$$\begin{aligned} Pr(\text{both good}) &= Pr(G_1 \cap G_2) \\ &= Pr(G_1)Pr(G_2|G_1) \\ &= [\qquad\qquad\qquad\qquad].\ (5/12) \end{aligned}$$

独立 (Independence)

Definition

事象 A は B と**独立 (statistically independent)** であるとは、

$$Pr(A|B) = Pr(A),$$

が成立することである。（事象 A が起こる確率は事象 B が起こった後に A が起こる確率と同じ。）すなわち、事象 A が B と独立である時、B という情報は A の確率に何の影響も与えない。

条件付確率の定義から、より一般に次のことが成り立つ。

Definition

$$Pr(A \cap B) = Pr(A)Pr(B)$$

が成り立つなら、A と B は**独立 (statistically independent)** である。

ベイズの定理 (Bayes' Theorem)

(例) あなたの上司は温厚な人で 99.9% の人とは上手に付き合っており、残り 0.1% の人とだけ仲が悪いとする。上司は仲の良い人には 1% の確率で返事が遅れ、仲の悪い人には 100% の確率で返事が遅れるとします。上司からの返事が遅いとき、その理由は上司があなたを嫌っているからである確率はどのくらいだろうか？

Theorem

ベイズの定理 *(Bayes' Theorem)*：互いに排反な事象 A_1, A_2 を原因事象とし、B を結果事象とする。このとき、結果 B が原因 A_1 によって生じたものである確率は、

$$Pr(A_1|B) = \frac{Pr(B|A_1)P(A_1)}{Pr(B)} = \frac{Pr(B|A_1)P(A_1)}{Pr(B|A_1)Pr(A_1) + Pr(B|A_2)Pr(A_2)}$$

となる。

3-23 The table below classifies the 115.5 million civilians in the 1985 U.S. labor force by age and employment status (Stat. Abst. of U.S.,1987, p.378):

	Age		
	Y (young, under 25)	O (older, 25 and over)	Totals
E (employed)	20.4	86.8	107.2
U (unemployed)	3.2	5.1	8.3
Totals	23.6	91.9	115.5 million

a. What is $Pr(U)$, the probability that a worker drawn at random will be unemployed? That is, find the unemployment rate.
b. What is $Pr(U|Y)$?
c. Is unemployment independent of age?

3-18: Suppose that 4 defective light bulbs inadvertently have been mixed up with 6 good ones.
a. If 2 bulbs are chosen at random, what is the chance that they both are good?
b. If the first 2 are good, what is the chance that the next 3 are good?
c. If we started all over again and chose 5 bulbs, what is the chance they all would be good?

3-35: Suppose that A and B are independent events, with $Pr(A) = .6$ and $Pr(B) = .2$. What is

a. $Pr(A|B)$?

b. $Pr(A \cap B)$?

c. $Pr(A \cup B)$?

3-36 Repeat Problem 3-35 if A and B are mutually exclusive instead of independent.

3-45: True or False? If false, correct it:
a. When two events are independent, the occurrence of one event will not change the probability of the second event.
b. Two events are mutually exclusive if they have no outcomes in common.
c. Two events are mutually exclusive if $Pr(A \cap B) = Pr(A)Pr(B)$.
d. If a fair coin has been fairly tossed 5 times and has come up tails each time, on the sixth toss the conditional probability of tails will be $1/64$.

5 章 確率変数と確率分布

東京大学

2014

離散確率変数 (Discrete Random Variables)

Example 2, continued: 一組の夫婦が 3 人の子供を産みたいと考えており、「女の子の数」に興味を持っているとする。この場合、

$$X = \text{女の子の数}$$

とすると、X は**確率変数 (random variable)** と呼ばれる。X の取り得る値は [　　] である。

例えば、女の子の産まれる確率が 48% であるとする。

確率変数 X の取り得る各値 x に対する確率を計算するためには、元の標本空間に戻る必要がある。

(例)

$$Pr(X = 1) = [\quad\quad] \ (.39)$$

同様にして、X の取り得る各値 x に対する確率は以下のように計算できる。

$$
\begin{aligned}
p(0) &\equiv Pr(X = 0) = [\qquad\qquad\qquad\qquad] \ (.14) \\
p(1) &\equiv Pr(X = 1) = [\qquad\qquad\qquad\qquad] \ (.39) \\
p(2) &\equiv Pr(X = 2) = [\qquad\qquad\qquad\qquad] \ (.36) \\
p(3) &\equiv Pr(X = 3) = [\qquad\qquad\qquad\qquad] \ (.11)
\end{aligned}
$$

以上の確率をまとめて **確率分布 (probability distribution)** と呼ぶ。

確率分布を用いることで、次のような問いに答えることもできる：「女の子が 2 人未満の確率は？」

$$
\begin{aligned}
Pr(X < 2) &= p(0) + p(1) \\
&= [\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad] \ (.53)
\end{aligned}
$$

期待値 (expectation) 、分散 (variance)、標準偏差 (standard deviation)

X の確率分布が分かると、事前に X の期待値、分散、標準偏差を求めることができる。

Definition

離散確率変数 X の取り得る値を $\{x_1, \ldots, x_m\}$ とする。

$$
\begin{aligned}
\textit{Mean } \mu &= E[X] = \sum_{i=1}^{m} x_i p(x_i) \\
\textit{Variance } \sigma^2 &= V(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i) \\
\textit{Standard Deviation } \sigma &= \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 p(x_i)}
\end{aligned}
$$

分散 σ^2 の計算のためには、次の便利な公式がある。

便利な公式

$$\sigma^2 = \sum_{i=1}^{m} x_i^2 p(x_i) - \mu^2$$

Example 2, continued: 前と同じ例において、X の期待値、分散、標準偏差はどうなるだろうか？(We suppose that the probability of a boy is 52%.)

$$X = 3\text{ 人の子供の中で女の子の数.}$$

$$\begin{aligned} \text{Mean } \mu &= [\quad\quad] \ (1.44) \\ \text{Variance } \sigma^2 &= [\quad\quad] \ (.75) \\ \text{Standard Deviation } \sigma &= [\quad\quad] \ (.87) \end{aligned}$$

二項分布 (Binomial Distribution)

離散確率変数には様々な種類のものがあるが、最も重要なものは**二項 (確率) 変数 (binomial variable)** と呼ばれる。

二項確率変数の古典的な例：

$$S = n \text{ 回の独立なコイン投げにおいて表が出る回数.}$$

一般に、ある試行において、特定の事象 A が起これば成功、それ以外であれば失敗とする。この時、成功回数 S は**二項 (確率) 変数 (binomial variable)** と呼ばれる。

Examples：

S ＝ 三人の子供のうちの女の子の数

S ＝ 選択式（ex. 5 択）の質問に当てずっぽうに答えた場合の正解数

S ＝ 選挙において無作為に選ばれた投票者中の共和党支持者の数

S ＝ 一日の工場生産において無作為に選ばれた製品の中の不良品の数

n 回の独立試行における「成功」数を S とする。各試行の成功確率を π とする時、n 回のうち s 回成功する確率は次のように表される。

二項分布の確率関数

$$p(s) = \binom{n}{s} \pi^s (1-\pi)^{n-s} \tag{1}$$

where

$$\begin{aligned} \binom{n}{s} &\equiv \frac{n!}{s!(n-s)!}, \text{ and} \\ n! &\equiv n(n-1)(n-2)\cdots 1 \end{aligned}$$

例えば、前の子供の性別の例で、女の子の数が 1 人となる確率は、
$n = [\quad]$, $\pi = [\qquad]$, and $s = [\quad]$ を (1) に代入して

$$p(1) = [\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad] \text{ (.39)}$$

となる。

Derivation of the binomial formula (1):
Let's consider a tossing coin $n = 5$ times. We suppose the coin is somewhat biased, coming up H with probability $\pi = .60$, and T with probability $1 - \pi = .40$. Here, let's calculate the probability that the number of H, $S = 3$. The generalization is straightforward.
One of the many ways we could get $S = 3$ is

$$HHH\ TT,$$

whose probability is

$$(.60)(.60)(.60)(.40)(.40) = (.60)^3(.40)^2.$$

But there are many other ways we could get exactly 3 heads in 5 tosses. For example, we might get the sequence ($THTHH$), which has a probability

$$(.40)(.60)(.40)(.60)(.60) = (.60)^3(.40)^2,$$

which is the same probability as before. In fact, all sequences in the event $S = 3$ will have this same probability.

Finally, how many ways are there to get exactly 3 heads? The answer is the number of of different ways that three H's and and two T's can be arranged. The number of arrangements of five distinct objects is

$$5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5!,$$

but we over-counted $(3 \cdot 2 \cdot 1)(2 \cdot 1)$ times. So the number of ways is

$$\binom{5}{3} = \frac{5!}{3!2!} = 10.$$

Hence we have

$$p(3) = \binom{5}{3}(.60)^3(.40)^2 = 10(.035) = .35$$

Example: How reliable is a small poll?
Suppose that a sample of 5 voters is to be randomly drawn from the U.S. population, when 60% vote Republican.
(a) The number of Republican voters in this sample of 5 can vary anywhere from 0 to 5. Tabulate its probability distribution.
(b) Calculate the mean and standard deviation of the number of Republican voters in this sample.
(c) What is the probability of exactly 3 Republican voters in the sample? (35%)
(d) Calculate the probability that the sample will have a majority of Republican voters (that is, at least 3) and thus will correctly reflect the population majority. (68%)

連続確率変数 (continuous random variables)

- 確率変数 X の取りうる値が連続的な場合、X は連続確率変数であるという。
- X が連続確率変数の場合、取りうる値は無数に存在するため、特定の点に有限の「確率」を付与すると全ての確率の和が無限になってしまう。そこで、確率は（点ではなく）幅に対して定義し、**密度関数** $f(x)$ を次のように定義する。

連続確率変数の場合の確率と密度関数

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

- 確率の和は 1 なので、次の事実が成り立つ。

$$Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1.$$

正規分布 (Normal Distribution)

正規分布は、次のような理由で統計において最も重要な分布である。

(i) Empirically, many random variables have bell-shaped probability densities which are very close to the normal distribution.
(ii) Errors made in measuring physical and economic phenomena often are normally distributed.
(iii) Many other probability distributions (such as the binomial) often can be approximated by the normal curve.
(iv) There is a famous theorem (Central Limit Theorem): The distribution of sum of ANY random variables approaches the normal distribution as the number of the observations increases (under some conditions). This fact will be very useful in statistical inference.

正規分布 (Normal Distribution)

以下の密度関数を持つ確率変数を**正規確率変数 (Normal random variable**、その分布を**正規分布 (Normal distribution)** という。

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

- 正規確率変数 X の期待値は μ、分散は σ^2 となる。
- X が正規分布に従うことを

$$X \sim N(\mu, \sigma^2)$$

と表す。

標準正規分布 (Standard Normal Distribution)

正規確率変数のうち、特に期待値 $\mu = \mathbf{0}$、分散 $\sigma^2 = \mathbf{1}$ であるものを、特に**標準正規確率変数 (Standard normal random variable)** といい、その分布を**標準正規分布 (Standard normal distribution)** という。

標準正規分布 (Standard normal distribution)

以下の密度関数を持つ確率変数を**正規確率変数**、その分布を**正規分布**という。

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

「標準正規確率変数 Z がある値以上 (あるいは以下）になる確率」に興味があるとする。

この確率は、統計家によって既に計算されており、標準正規分布表にまとめられている。

この表を使って以下のような確率も計算することができる。

Example: If Z has a standard normal distribution, find:
a. $Pr(Z > 1.64)$
b. $Pr(Z < -1.64)$
c. $Pr(1.0 < Z < 1.5)$
d. $Pr(-1 < Z < 2)$
e. $Pr(-2 < Z < 2)$

一般の正規分布 (General Normal Distribution)

In general, a normal distribution may have any mean μ, and any standard deviation σ. For example, when the population of American men have their height X arrayed into a frequency distribution, it looks a normal distribution with mean $\mu = 69$ inches, and standard deviation $\sigma = 3$ inches.

Question: How could we calculate the proportion of men above 74 inches, for example?

The important fact we use is

$Z = \frac{X-\mu}{\sigma}$ **follows the standard normal distribution** when X is normally distributed with mean μ and standard deviation σ.

Since we have the table of tail probabilities of the standard normal distribution (the standard normal table), we can calculate

$$\begin{aligned} Pr(X > 74) &= Pr(\frac{X-\mu}{\sigma} > \frac{74-\mu}{\sigma}) \\ &= Pr(Z > 1.67) \approx 5\% \quad \text{(we use the stander normal table here} \end{aligned}$$

Sometimes, we say that

"74 inches is $\frac{74-\mu}{\sigma} = 1.67$ standard deviation away from the mean."

4-20: If X is normally distributed around a mean of 16 with a standard deviation of 5, find

a. $Pr(X > 20)$

b. $Pr(X < 10)$

c. $Pr(20 < X < 25)$

d. $Pr(12 < X < 24)$

確率変数の関数の期待値

Since means play a key role in statistics, they have been calculated by all sorts of people, who sometimes use different names for the same concept. For example, geographers use the term "mean annual rainfall," teachers use the term "average grade," and gamblers and economists use the term "expected profit."
We often use the the following notation E called "expected value", and define in general

Definition

$$E[g(X)] = \sum_x g(x)p(x)$$

for the mean of a function of $g(X)$.

As an example, one possible form of the function $g(x)$ is:

$$g(X) = (X - \mu)^2.$$

Then we have

$$E[(X - \mu)^2] = \sum_x (x - \mu)^2 p(x),$$

which equals to the variance of X. Hence we can write

$$\text{Variance } \sigma^2 = E[(X - \mu)^2].$$

In the new E notation, we also can rewrite (2) as

$$\sigma^2 = E(X^2) - \mu^2$$

4-10: In families with 6 children, let $X =$ the number of boys. For simplicity, assume that births are independent and boys and girls are equally likely.

a. Graph the probability distribution of X.

b. Calculate the mean and standard deviation, and show them on the graph.

c. Of all families with 6 children, what proportion have:

i. Exactly an even split between the sexes (3-3)?

ii. Nearly an even split (3-3 or 4-2 or 2-4)?

iii. 3 or more boys?

4-22: The time required to complete a college achievement test was found to be normally distributed, with a mean of 110 minutes and a standard deviation of 20 minutes.

a. What proportion of the students will finish in 2 hours (120 minutes)?

b. When should the test be terminated to allow just enough time for 90% of the students to complete the test?

4-25: In her new job of selling computers, Dawn Elliot faces uncertain prospects next year. She guesses that her taxable income X might be anywhere from 20 to 50 thousand dollars according to her schedule of personal probabilities $p(x)$ given below. The corresponding tax is given in the final column.

Income x (\$000)	Probability $p(x)$	Tax $t(x)$ (\$000)
20	0.1	4
30	0.3	6
40	0.4	9
50	0.2	13

a. Calculate her expected income.

b. Calculate her expected tax.

c. Calculate her expected disposable income (after tax) in two ways:

i. Calculate first the table of disposable incomes, and then take their expected value.

ii. As an easier way, just use the answers in **a** and **b**.

複数の確率変数：同時確率分布 (Joint Distributions)

There are many cases we want to consider two random variables at the same time. In the planning of three children, let us denote two random variables:

$$\begin{aligned} X &= \text{number of girls} \\ Y &= \text{number of runs} \end{aligned}$$

where a *run* is an unbroken string of children of the same sex. For example, $Y = 1$ for the outcome BBB, while $Y = 2$ for the outcome BBG.

Suppose that we are interested in the probability that a family would have 1 girl and 2 runs, $Pr(X = 1 \text{ and } Y = 2)$? (We assume that the probability of a boy is 52%.)

$$Pr(X = 1 \text{ and } Y = 2) = [\qquad\qquad]$$

which we simply denote by $p(1, 2)$.
Similarly, we could compute $p(0, 1)$, $p(0, 2)$, $p(0, 3)$, $p(1, 2), \ldots$, obtaining the **joint probability distribution** of X and Y.

x \ y	1	2	3
0	0.14	0	0
1	0	0.26	0.13
2	0	0.24	0.12
3	0.11	0	0

Table : Joint Distribution $p(x, y)$

周辺確率分布 (Marginal Distributions)

Suppose that we are interested only in X, yet have to work with the joint distribution of X and Y. For any given x, we define the **marginal distribution** of x:

X の周辺分布

$$p_X(x) = \sum_y p(x, y).$$

For example,

$$p_X(2) = p(2,1) + p(2,2) + p(2,3) = \sum_y p(2, y)$$

Of course, the distribution of Y can be calculated in a similar way:

Y の周辺分布

$$p_Y(y) = \sum_x p(x, y).$$

独立 (Independence)

Definition

X and Y are independent if

$$p(x, y) = p(x)p(y)$$

for all x and y.

For the distribution in Table 1, are X and Y independent?
Answer: [ ]

複数の確率変数の関数の期待値

Suppose we are interested in the expected value of some function of X and Y, $g(x, y)$. We define it

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)p(x, y),$$

which is similar to the earlier formula

$$E[g(X)] = \sum_x g(x)p(x)$$

in the one variable case.

共分散と相関係数

- One of the most interesting questions in the face of two variables are how they vary together or how they are related. In this section we will develop how they can be measured. We define the **covariance** first, which is given by

$$\sigma_{XY} = \text{Covariance of } X \text{ and } Y \equiv E[(X - \mu_X)(Y - \mu_Y)]$$

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$.

- The calculation of the covariance can often be simplified by using an alternative formula:

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y$$

$\sigma_{XY} > 0$ indicates the positive relation between X and Y (when one is large, the other tends to be [ ].)

$\sigma_{XY} < 0$ indicates the negative relation between X and Y (when one is large, the other tends to be [ ].)

If X and Y are independent, then they are uncorrelated ($\sigma_{XY} = 0$).

Since the covariance depends on the units in which X and Y are measured, [1] it is not a good measure of the strength of the relation of the two variable. We define the correlation ρ which is completely independent of the scale in which either X and Y is measured.

$$\text{Correlation, } \rho \equiv \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

[1] If X, for example, were measured by "*cm*" instead of "*m*", σ_{XY} would unfortunately increase by 100 times.

Now suppose that X and Y have a perfect positive linear relation. For example, suppose they always take on the same values, so that $X = Y$. Then it is easy to show that ρ takes on [ ].
Similarly, if there is a perfect negative linear relation, then ρ would be [ ].
In fact, ρ is always bounded:

$$-1 \leq \rho \leq 1.$$

5-15: Does education make you happy?

Americans were asked in 1971 to rank themselves on a "happiness index" as follows: $H = 0$ (not happy), $H = 1$ (fairly happy), or $H = 2$ (very happy). The amount of education was also measured for each individual: $X = 1$ (elementary school completed), $X = 2$ (high school completed), or $X = 3$ (college completed, or more). Thus $X =$ number of schools completed. Then the relative frequencies of various combinations were roughly as follows (Gallup, 1971):

a: Calculate the covariance and correlation.

	h		
x	0	1	2
1	0.02	0.08	0.05
2	0.02	0.28	0.25
3	0.01	0.13	0.16

Table : Joint Distribution of Happiness H and Education X

b: Answer True or False; if False, correct it.

i. As X increases, the average level of H increases. This positive relation is reflected in a positive correlation ρ.

ii. Yet the relation is just a tendency (H fluctuate around its average level), so that ρ is less than 1.

iii. This shows that those who receive high education tend to be happier than the poor, that is, education tends to make people happier.

Linear combination of Two Random Variables

$$E[X + Y] = E[X] + E[Y]$$

$$E[aX + bY] = aE[X] + bE[Y]$$

for any constants a and b.

Proof:

$$
\begin{aligned}
E[aX + bY] &= \sum_x \sum_y (ax + by) p(x, y) \\
&= a \sum_x x (\sum_y p(x, y)) + b \sum_y y (\sum_x p(x, y)) \\
&= a \sum_x x p(x) + b \sum_y y p(y) \\
&= aE(X) + bE(Y)
\end{aligned}
$$

Variance

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\ Cov(X,Y)$$

for any constants a and b.

Proof:

$$
\begin{aligned}
& Var(aX + bY) \\
= & \sum_x \sum_y \{(ax + by) - (a\mu_x + b\mu_y)\}^2 p(x, y) \\
= & \sum_x \sum_y \{a(x - \mu_x) + b(y - \mu_y)\}^2 p(x, y) \\
= & \sum_x \sum_y \{a^2(x - \mu_x)^2 + b^2(y - \mu_y)^2 + 2ab(x - \mu_x)(y - \mu_y)\} p(x, y) \\
= & a^2 \sum_x \sum_y (x - \mu_x)^2 p(x, y) + b^2 \sum_x \sum_y (y - \mu_y)^2 p(x, y) \\
& + 2ab \sum_x \sum_y (x - \mu_x)(y - \mu_y) p(x, y) \\
= & a^2 Var(X) + b^2 Var(Y) + 2ab\ Cov(X, Y)
\end{aligned}
$$

Example: Investors often prefer to have a portfolio whose expected profit is as high as possible and variance is as low as possible. How can we construct such a "nice" portfolio?
Suppose there are two assets X and Y with $E[X] = 5$, $E[Y] = 5$, $Var(X) = 1$, $Var(Y) = 1$. What are the expected profit and the variance of the portfolio $Z = \frac{1}{2}X + \frac{1}{2}Y$ in the following three cases?
a. When $Cov(X, Y) = 0.3$.
b. When $Cov(X, Y) = -0.3$.
c. When X and Y are independent.

If X and Y are **independent**,

$$Var(X + Y) = Var(X) + Var(Y)$$

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

for any constants a and b.