第2章：確率と統計

Jonathan Roth

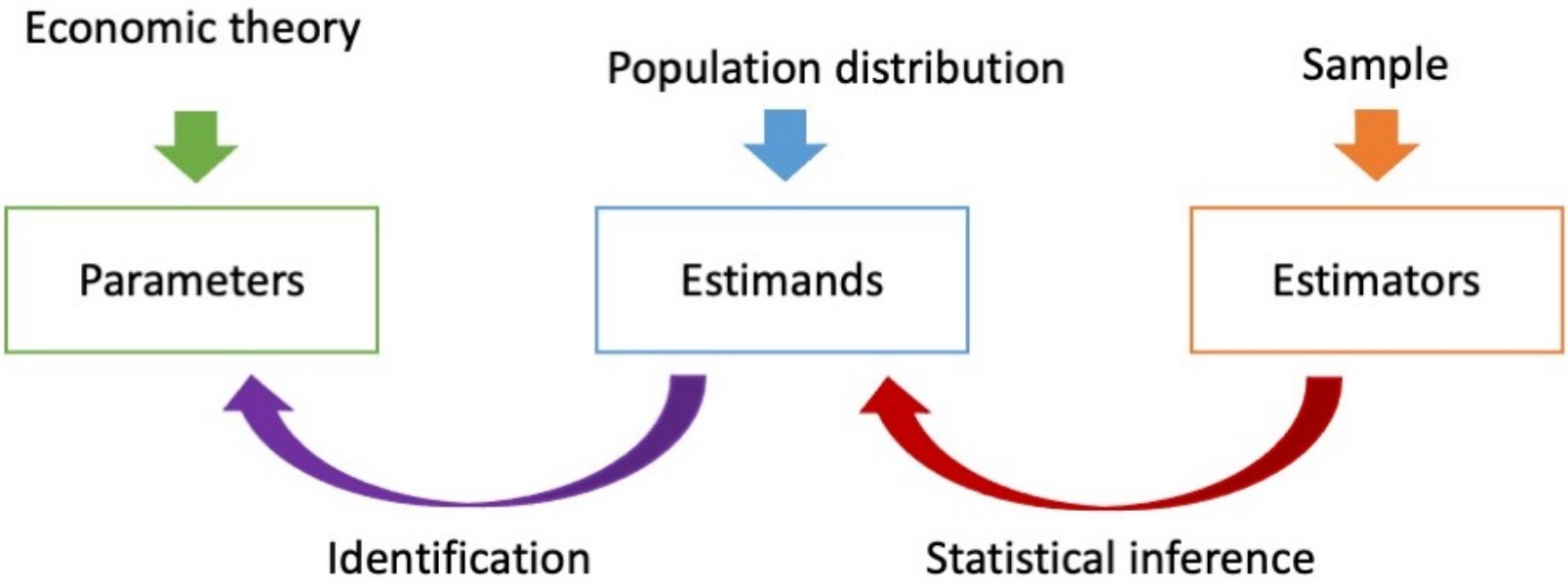
数理計量経済学 I
ブラウン大学

連絡事項

- 課題（Problem Set）1 を公開しました。締め切りは 9 月 19 日（金）午後 4 時、Gradescope での提出です。
- 今週から TA セッションとオフィスアワー（OH）が始まります。時間と暫定的な場所については Canvas を確認してください。
- 運営面で質問はありますか？

「全体像（Big Picture）」の復習

役割分担を思い出しましょう：



- 統計（Statistics）：観察されたサンプルデータは、関心のある母集団の観察可能な特徴とどう関連しているか？
- 識別（Identification）：母集団の観察可能な特徴は、関心のあるターゲット・パラメータとどう関連しているか？
- これら両方のタスクにおいて、データがどのように生成されるかを議論するための数学的な言語が必要です。それが確率と統計です。


I love
Statis - sticks

アウトライン

1. 確率変数と確率分布
2. 平均と分散
3. 実験における識別
4. 無作為抽出と標本平均
5. 仮説検定と推論

確率変数

- 確率論は、ランダムな過程（random processes）の研究を定式化したものです。

- ランダムな過程の例には何があるでしょうか？
  - コイン投げ – 表か裏か？
  - 米国の世帯を無作為に調査 – 所得はいくらか？

- ランダムな過程の結果（実現値）を確率変数（random variable）と呼びます。

用語の整理

- 結果（**Outcomes**）とは、ランダムな過程における互いに排他的な結果のことです（例：1 回のコイン投げの結果は「表」と「裏」）。
- 確率（**Probability**）とは、ある結果が起こる尤もしさを捉えたものです（すなわち、過程を繰り返したときの発生頻度）。
- 標本空間（**Sample space**）とは、起こりうるすべての結果の集合です。
- 事象（**Event**）とは、標本空間の部分集合です。その確率は、含まれる結果の確率の和になります。

例：2 枚の公正なコインを投げます。

- 起こりうる結果（標本空間）は何ですか？
- 少なくとも 1 枚が表になる（事象）確率はいくらですか？

確率変数と累積分布関数 (CDF)

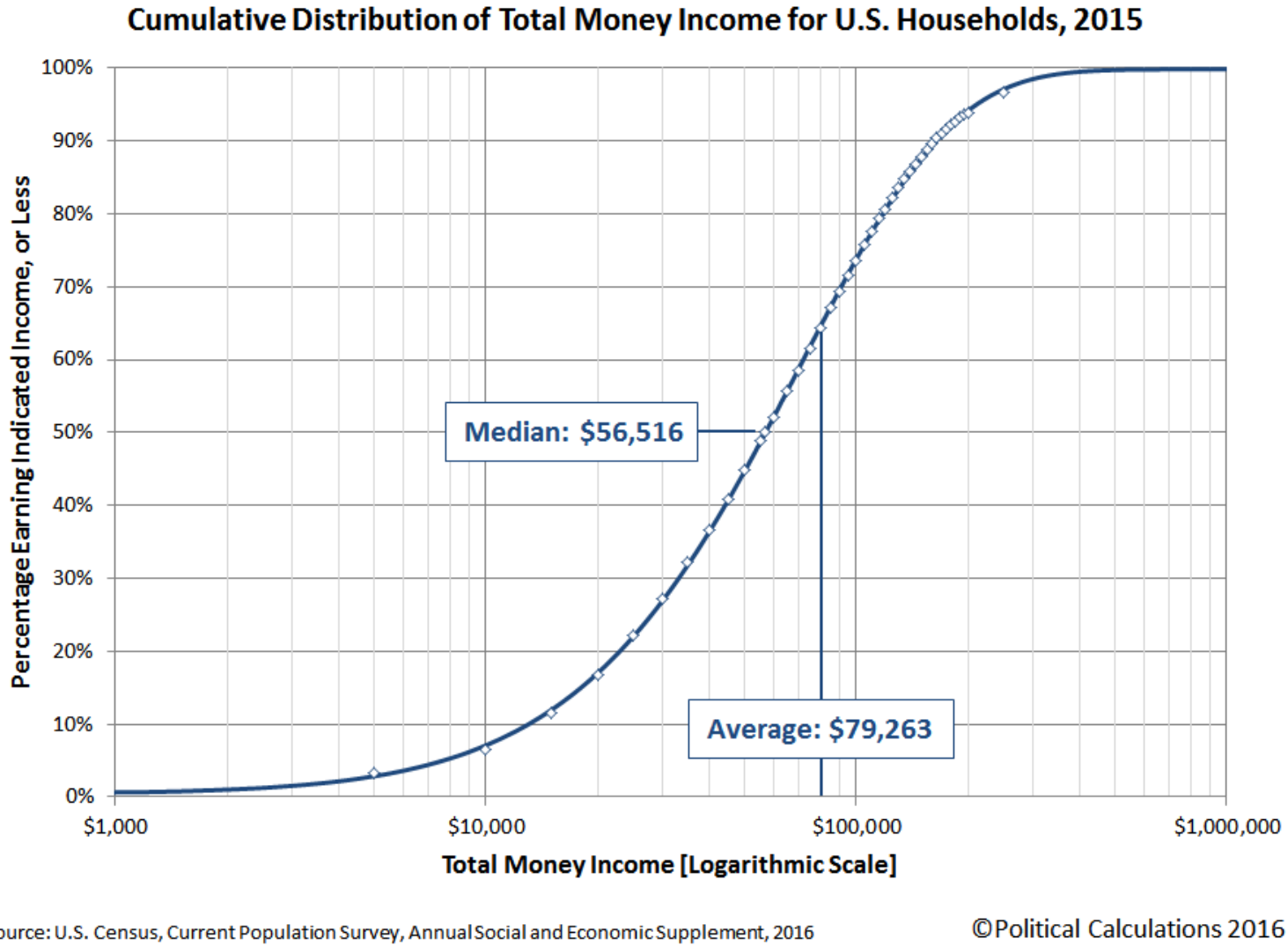
- 確率変数とは、ランダムな過程を数値で要約したものです（形式的には、標本空間上で定義された実数値関数）。
  - 例：X を表が出た回数とする。

- 実数値の確率変数は、その累積分布関数（**CDF**）によって特徴付けられます。

$$F(x) = Pr(X \leq x)$$

これは、X がある値 x 以下になる確率を示します。

  - 注意：通常、「X」のような確率変数の実現値（すなわちランダムではない数値）を表すには、「x」のような小文字を使用します。



- 2016 年、米国の世帯の約半分は所得が 5.6 万ドル以下でした。
- 形式的には：$F(56,516) = 0.5$

サポート（Support）

- 確率変数 X のサポート（$\mathbb{X}$ と表記）とは、X が取りうる値の集合です。
  - X が 1 年間のうち雇用されていた月数なら、$\mathbb{X} = \{0, 1, ..., 12\}$ です。
  - X が所得なら、$\mathbb{X} = \mathbb{R}_{\geq 0}$（近似的に）です。

- X のサポートが有限（例：$\{0, 1\}$）なら、その X は離散型であると言います。

- X のサポートが連続体（例：$\mathbb{R}$ や $[0, 1]$）なら、その X は連続型の分布に従うと言います（厳密には CDF が微分可能な場合）。

密度関数と質量関数

- X が離散型の場合、サポートの各値を取る確率として確率質量関数（PMF）を定義します：

$$p(x) = Pr(X = x)$$

  - 離散型確率変数の CDF は次のようになります：

$$F(x) = \sum_{x' \leq x} p(x')$$

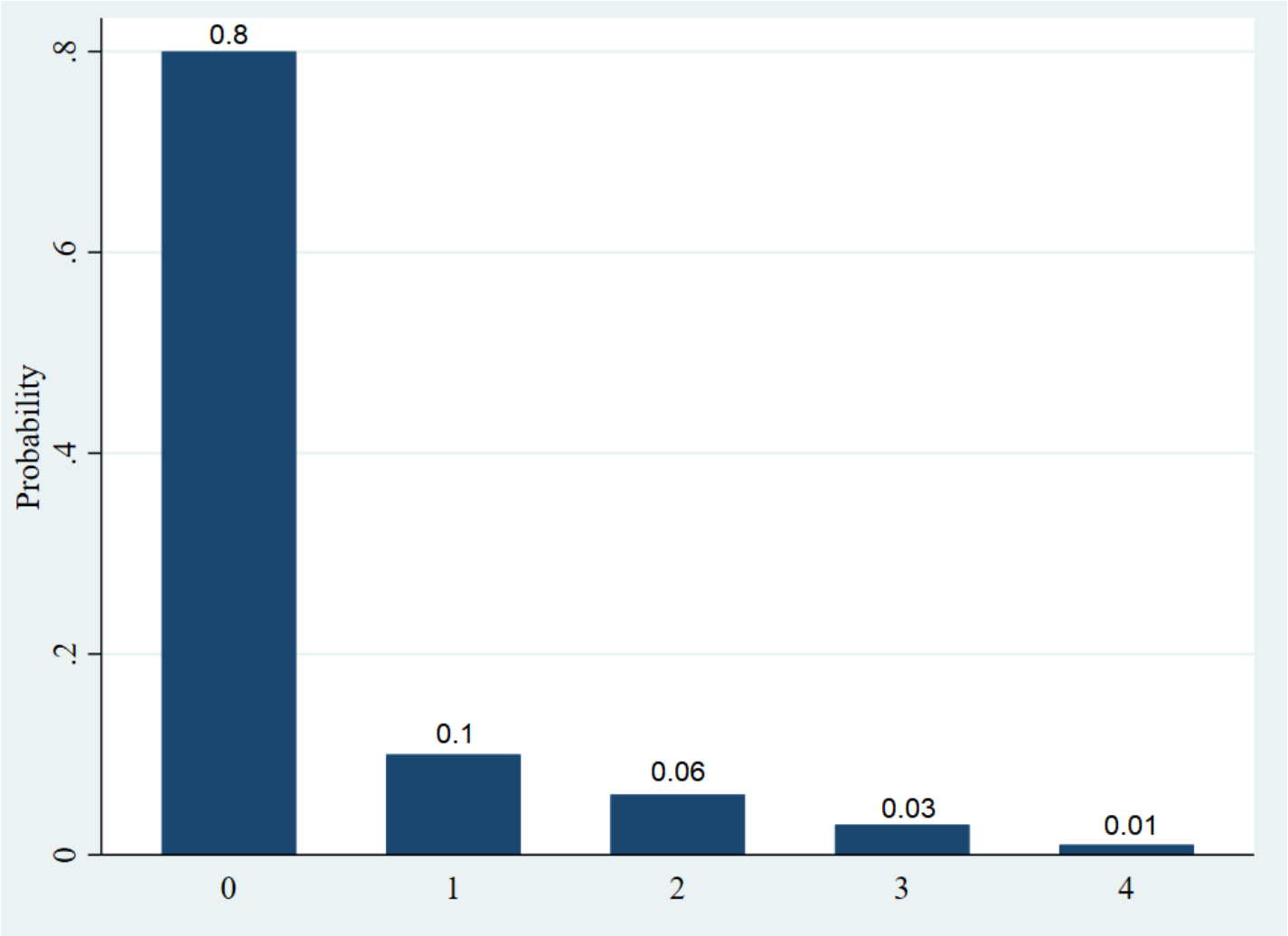
- 連続型確率変数の場合、確率密度関数（PDF）を $f(x) = \frac{d}{dx} F(x)$ と定義します。これは CDF が次であることを意味します：

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- 表記上の注意：$p(x)$ と $f(x)$ の両方が PDF/PMF として使われます。

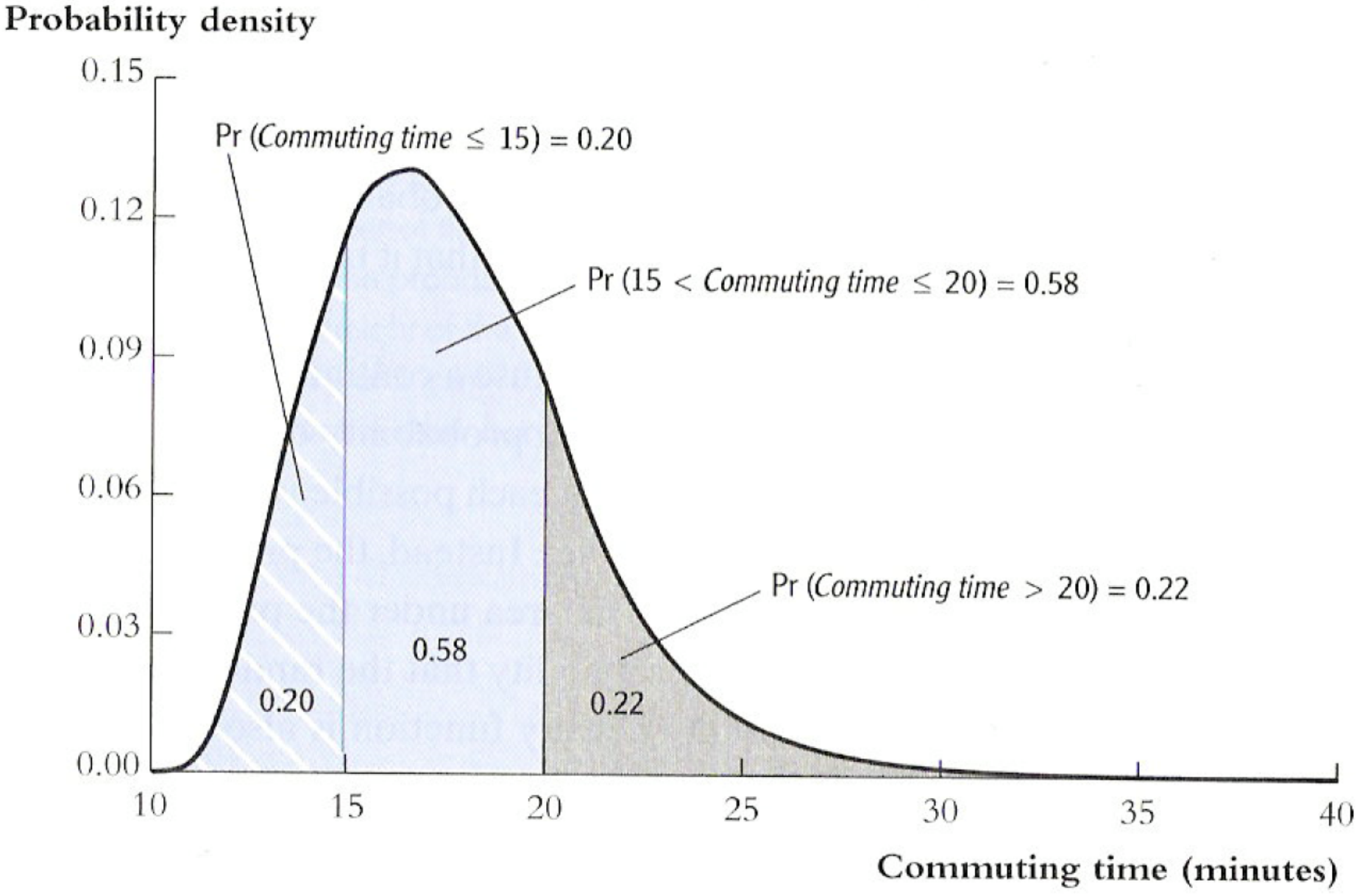
離散型確率変数の例：Wi-Fi 接続の失敗回数

- 以下は PMF です。CDF はどうなりますか？



連続型確率変数の例：通勤時間

- 以下は PDF です。CDF はどうなりますか？



PDF/CDF の性質

- CDF $F(x) = Pr(X \leq x)$ の主要な性質：
  - 非減少：$x > x'$ ならば $F(x) \geq F(x')$
  - $\lim_{x \to -\infty} F(x) = 0$ および $\lim_{x \to +\infty} F(x) = 1$ を満たす
- PDF $f(x) = \frac{\partial}{\partial x} F(x)$ の対応する性質：
  - 非負：すべての $x \in \mathbb{X}$ に対して $f(x) \geq 0$
  - $\int_{x \in \mathbb{X}} f(x) dx = 1$ を満たす
  - PMF の場合：$\sum_{x \in \mathbb{X}} p(x) = 1$

ベルヌーイ分布

重要な離散分布：ベルヌーイ分布 $X \in \{0, 1\}$

- 例：大学卒業の有無、コインが「表」になるかどうか（「ダミー変数」とも呼ばれます）
- PMF：

$$p(x) = \begin{cases} 1 - \pi, & x = 0 \\ \pi, & x = 1 \end{cases}$$

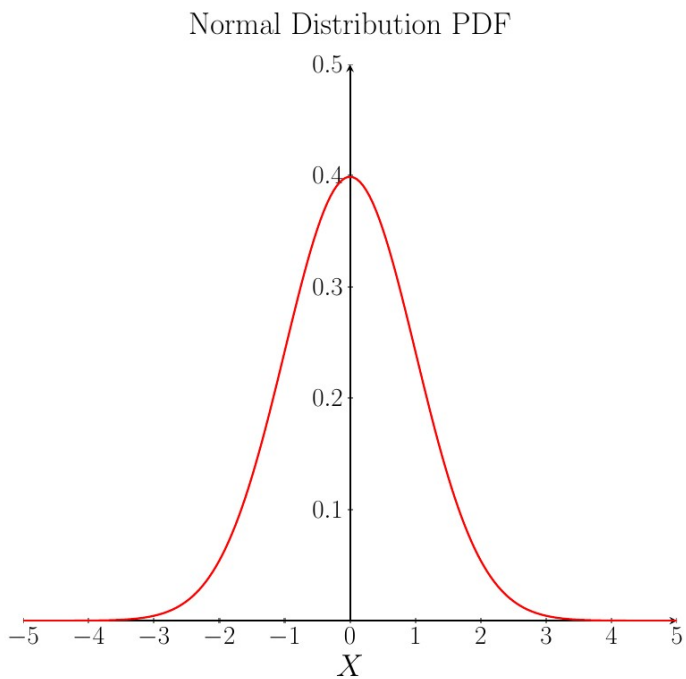
（ただし $\pi \in [0, 1]$）

- π は X の平均（期待値）です。これについては後ほど正式に定義します。
- $X \sim Bernoulli(\pi)$ と表記されます。
- 2 値変数 X の分布は、必ずこの形式になります。

正規分布と一様分布

重要な連続分布：正規分布 $X \in \mathbb{R}$

- 例：年収の対数（近似的に）
- PDF：$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$（ただし $x \in \mathbb{R}, \sigma > 0$）
- ここで μ は X の平均、σ^2 は分散です。
- $X \sim \mathrm{N}(\mu, \sigma^2)$ と表記されます。
- 有用な性質：X が正規分布に従うなら、任意の定数 a, b に対して $aX + b$ も正規分布に従います。



同時分布

経済学の興味深い問いの多くは、2 つ以上の確率変数の同時分布（joint distribution）の特徴に関わっています。

- 例：所得（Y）と教育水準（X）はどのように共変動するか？
- 同時分布の CDF は次のように定義されます：

$$F(x,y) = Pr(X \leq x, Y \leq y)$$

すなわち、$X \leq x$ かつ $Y \leq y$ となる確率です。

- 離散型 (X,Y) の場合、同時 PMF $p(x,y) = Pr(X = x, Y = y)$ を定義します。
- 連続型 (X,Y) の場合、同時 PDF $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$ を定義します。

複数の確率変数を扱う際、個々の変数の分布を周辺分布（marginal distribution）と呼ぶことがあります。

- 同時分布とは、例えば $p(x) = \sum_{y \in \mathbb{Y}} p(x,y)$ のような関係で結ばれています。

条件付き分布

同時分布と周辺分布を組み合わせることで、ある確率変数が与えられたときの別の確率変数の条件付き分布（**conditional distribution**）が得られます。

- 直感的には、条件付き分布 $Y|X=x$ とは、$X=x$ であるサブグループにおける Y の分布のことです。
- 条件付き PMF $p(y \mid x) = Pr(Y = y \mid X = x) = \frac{Pr(Y=y, X=x)}{Pr(X=x)} = \frac{p(y,x)}{p(x)}$
- 例：大学卒業を条件とした所得の分布。

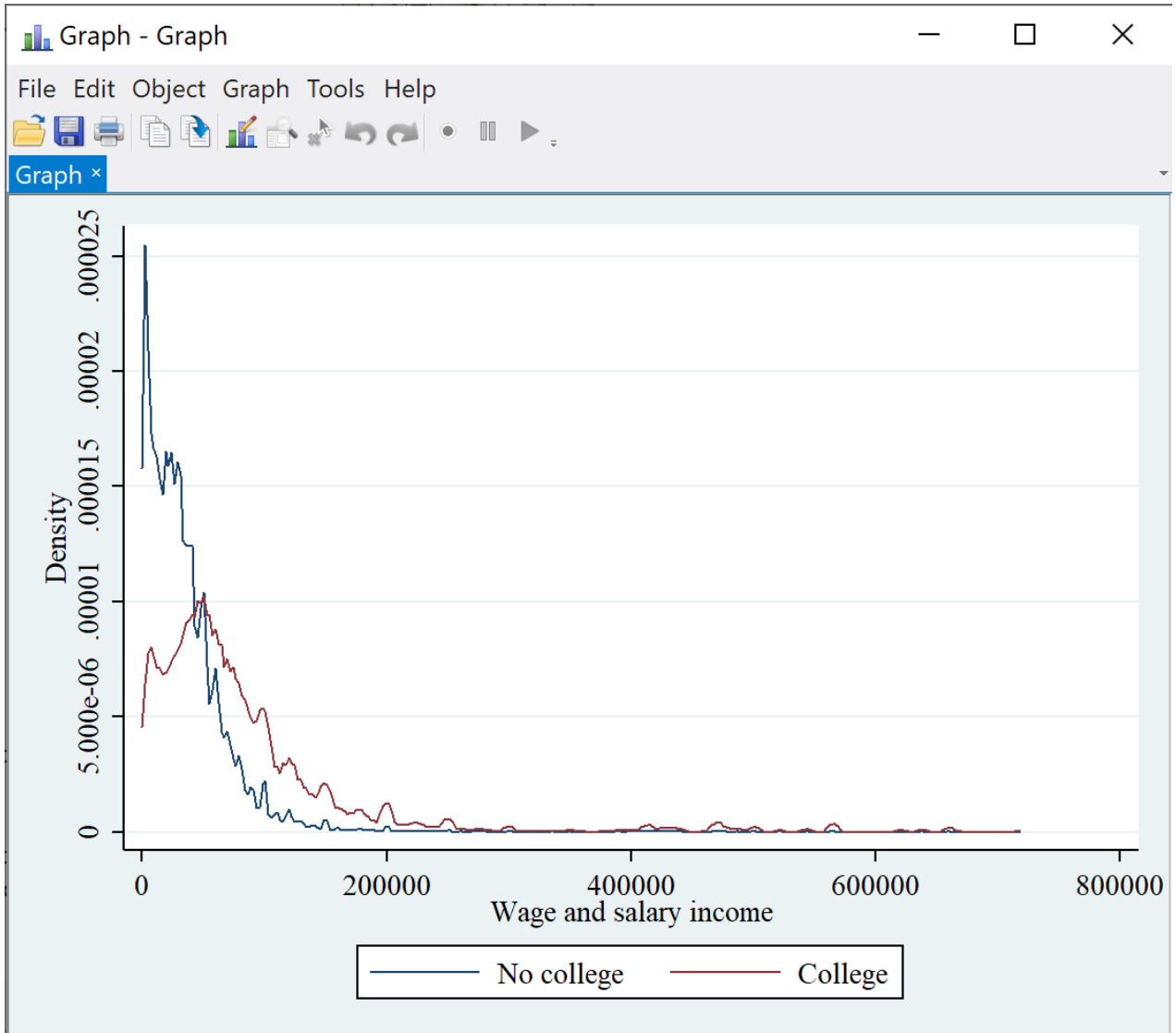
これは直ちにベイズの定理を導きます：

$$p(y \mid x) = p(x \mid y)\frac{p(y)}{p(x)}$$

確率変数と確率分布（続き）

大学卒業を条件とした年収の条件付き密度関数：

```
1  twoway (kdensity incwage if educ<10) (kdensity incwage if educ>=10), ///
2      xtitle("Wage and salary income") ytitle("Density") ///
3      legend(label(1 "No college") label(2 "College"))
```



独立性

- このコースで重要な概念となるのが独立性（**independence**）です。
- 直感的には、独立性とは、X の値を知っても Y の値について何も分からないということです。
- 形式的には、すべての (y, x) に対して条件付き分布が周辺分布と一致する場合、X と Y は独立である（$X \perp\!\!\!\perp Y$）と言います：

$$p(y \mid x) = p(y)$$

- 例：D がランダムに割り当てられた処置であるなら、$D \perp\!\!\!\perp (Y(1), Y(0))$ です。
  - 理解度チェック：これは $D \perp\!\!\!\perp Y$ を意味しますか？
- 条件付き独立性も同様に定義されます。W が与えられた下で X を知っても Y についての情報が増えない場合、$Y \perp\!\!\!\perp X \mid W$ と書きます：

$$p(y \mid x, w) = p(y|w)$$

多変量正規分布

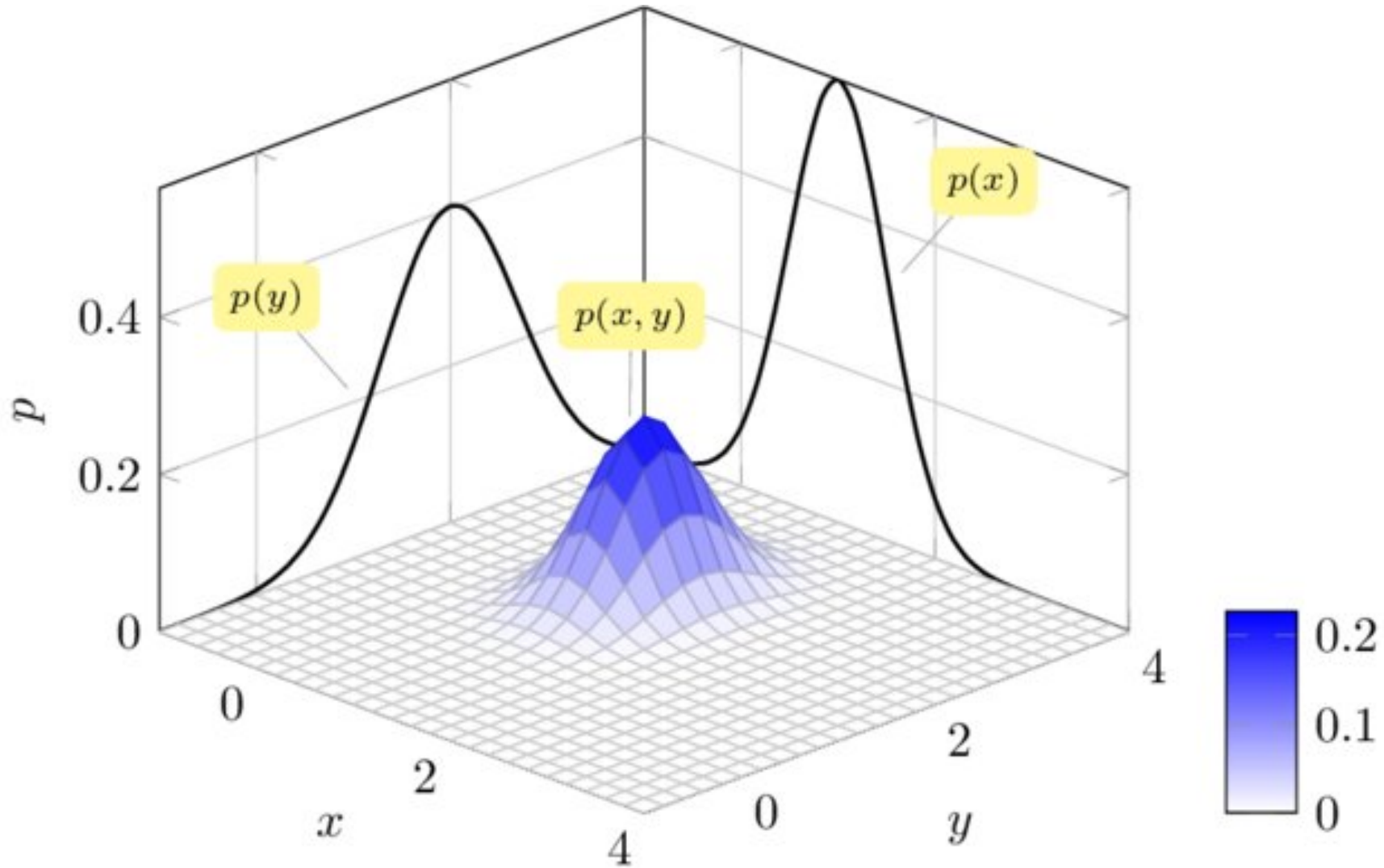
重要な多変量分布：多変量正規分布 $\mathbf{X} \in \mathbb{R}^K$

- 平均ベクトル $\boldsymbol{\mu} \in \mathbb{R}^K$ と、正定値な分散共分散行列 $\boldsymbol{\Sigma} \in \mathbb{R}^K \times \mathbb{R}^K$ によってパラメータ化されます。
- 注意：原則として、ベクトルや行列は太字で表記します。

多くの有用な性質がありますが、ここではいくつか紹介します。$(X, Y)'$ が多変量正規分布に従うとき：

- X と Y の周辺分布は正規分布になります。
- $X \mid Y$ および $Y \mid X$ の条件付き分布も正規分布になります。
- 任意の線形結合 $aX + bY + c$ も正規分布に従います。

2変量正規分布の密度関数



アウトライン

1. 確率変数と確率分布 ✓
2. 平均と分散
3. 実験における識別
4. 無作為抽出と標本平均
5. 仮説検定と推論

平均

- 私たちはしばしば、経済的な確率変数の平均（例：世帯所得）に興味を持ちます。

- X の平均（**mean**）/期待値（**expectation**）とは、確率で重み付けされた代表的な値のことです。
  - 離散型確率変数の場合：

$$E[X] = \sum_{x \in \mathbb{X}} p(x)x = x_1 Pr(X = x_1) + \cdots + x_K Pr(X = x_K)$$

    解釈：抽出を何度も繰り返したときの X の長期的な平均値。

  - 連続型確率変数の場合：$E[X] = \int_{x \in \mathbb{X}} f(x)xdx$
    注意：$p(x)$ が極端な値に対して高い確率を持つ場合、期待値が存在しないこともあります。

- 重要な事実：期待値演算子は線形です。定数 a, b に対して $E[a + bX] = a + bE[X]$。
  - これは上記の定義から容易に証明できます。

平均の計算：簡単な例

- X を公正なサイコロの目とします。$E[X]$ はいくらですか？
- 定義より、

$$E[X] = Pr(X = 1) \times 1 + Pr(X = 2) \times 2 + ... + Pr(X = 6) \times 6$$

- 公正なサイコロなら、$Pr(X = 1) = ... = Pr(X = 6) = \frac{1}{6}$ です。
- これを代入すると：

$$E[X] = \frac{1}{6}(1 + ... + 6) = 3.5$$

分散

分散（Variance）は、分布の広がりの二乗を測定します：

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E\left[X\right]^2$（なぜこうなるか分かりますか？）
- X の標準偏差（standard deviation）は $Std(X) = \sqrt{Var(X)}$ です。これは平均からの「典型的な」ズレを表します。

線形変換の分散：

$$\begin{aligned} Var(a + bX) &= E[(a + bX - E[a + bX])^2] \\ &= E[(a + bX - a - bE[X])^2] \\ &= b^2 E[(X - E[X])^2] \\ &= b^2 Var(X) \end{aligned}$$

これは $Std(a + bX) = |b| \cdot Std(X)$ であることを意味します。

- 直感的には、所得をドル単位からセント単位に変えると、標準偏差は 100 倍になります。

共分散

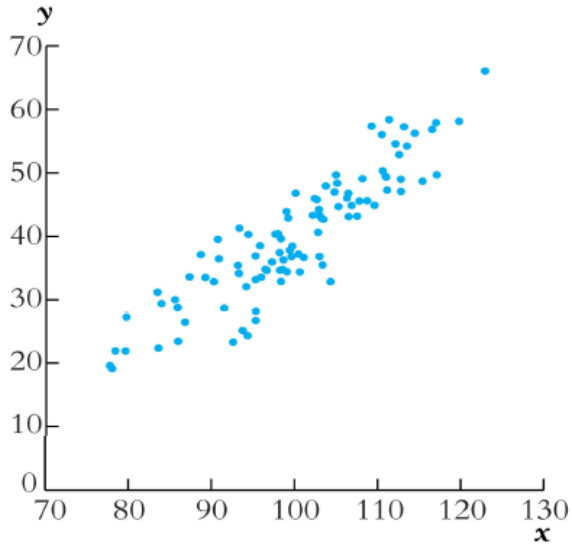
共分散（**Covariance**）は、2 つの変数の間の線形な関連性を測定します：

- $Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- $Cov(X, X) = E\left[(X - E[X])^2\right] = Var(X)$ となることに注目してください。
- X と Y の相関（**correlation**）は $Corr(X, Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$ です。

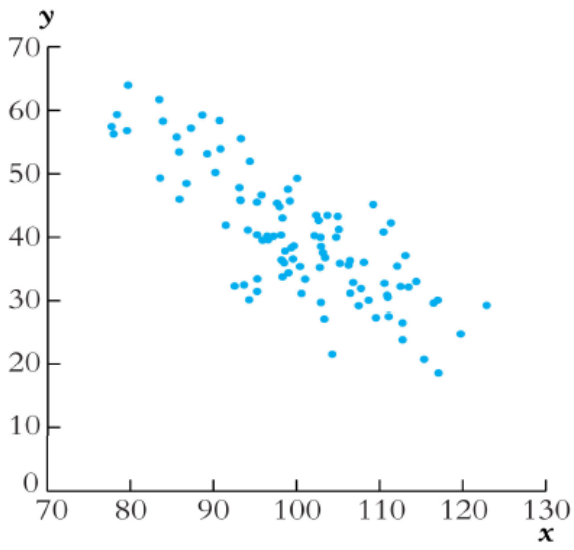
$Cov(X, Y) > 0$ は、X が平均より高いときに Y も平均より高くなる傾向があることを意味します（逆も同様）。

- $Corr(X, Y)$ は、線形な関連性を表す単位のない（標準化された）尺度です。
- X と Y が独立なら、$Cov(X, Y) = Corr(X, Y) = 0$ です。
- しかし、逆は必ずしも成り立ちません！ 独立性は相関ゼロよりも強い概念です。

相関の例

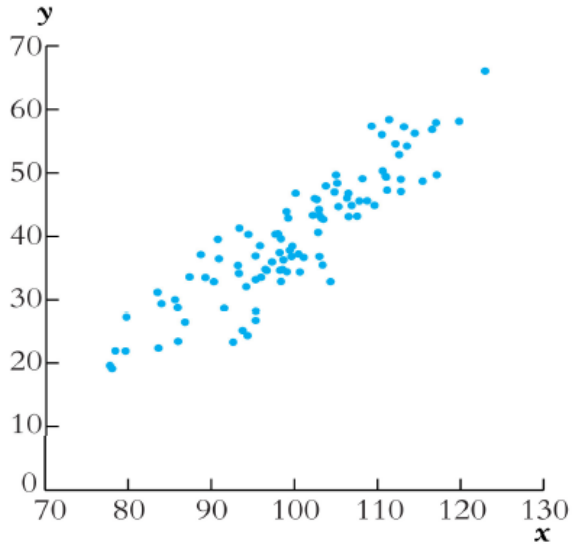


(a) Correlation = +0.9

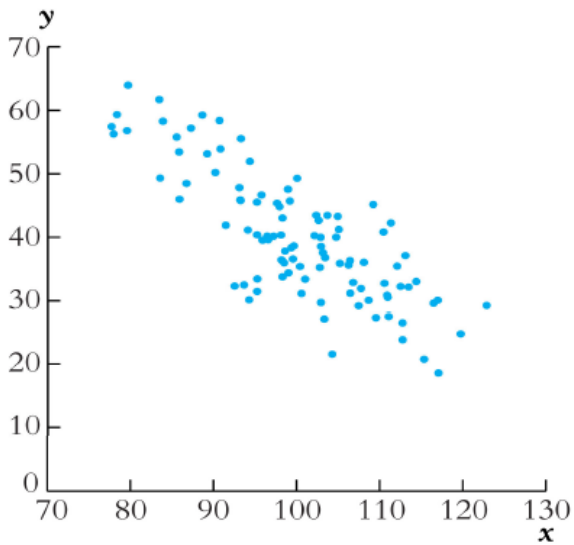


(b) Correlation = –0.8

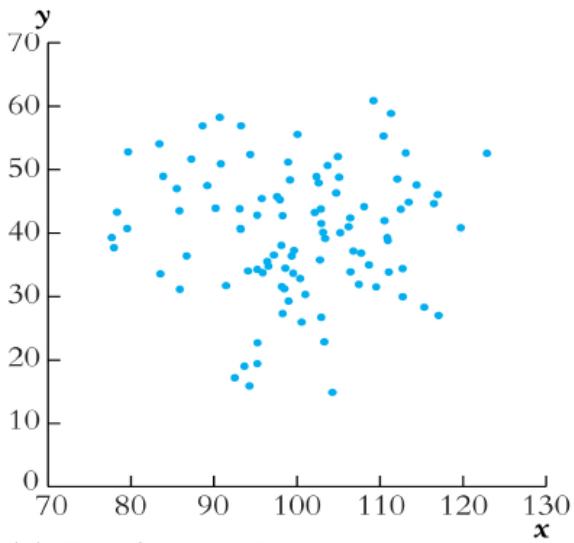
相関の例



(a) Correlation = +0.9

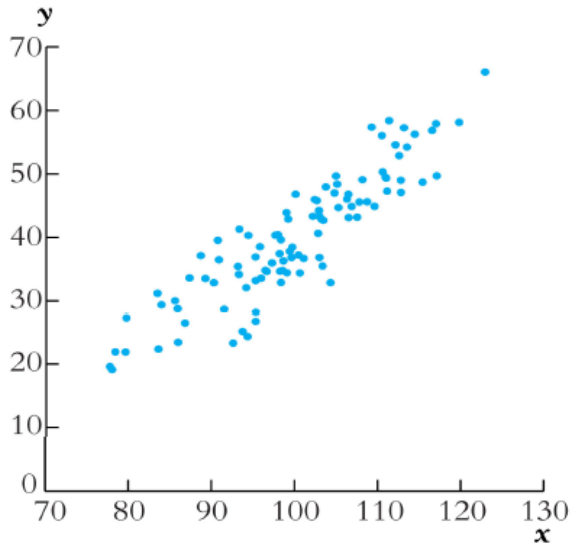


(b) Correlation = –0.8

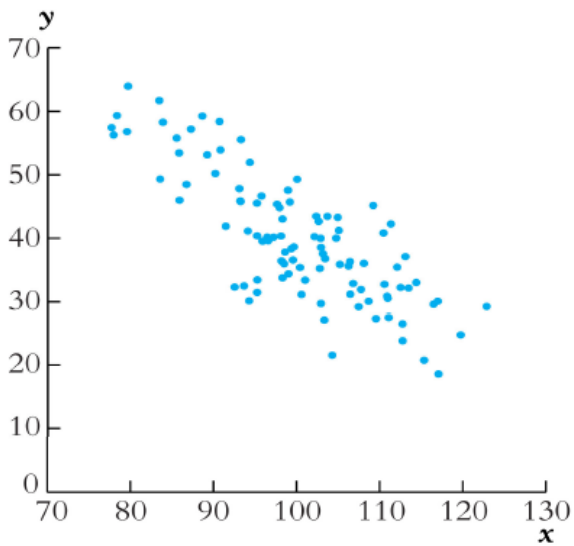


(c) Correlation = 0.0

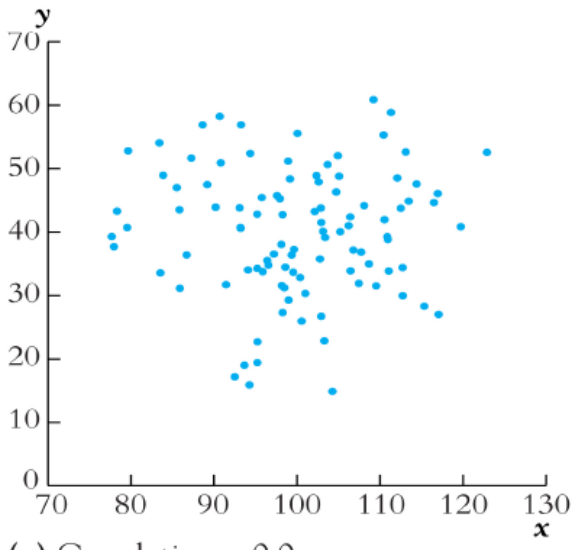
相関の例



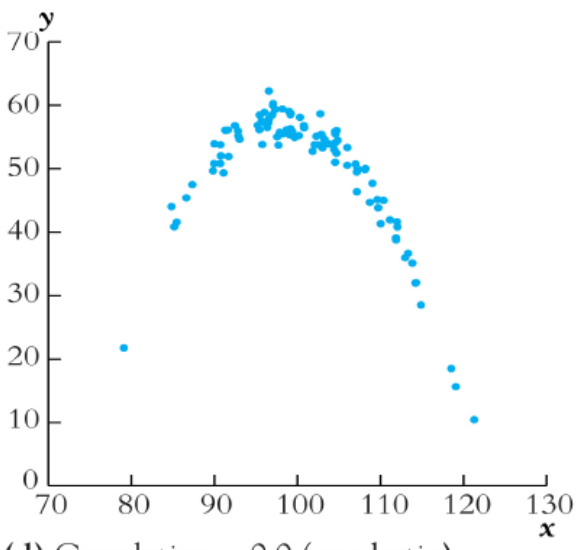
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

線形結合の平均と分散

- **期待値は線形です：** $E[aX + bY + c] = aE[X] + bE[Y] + c$
- **分散は「2 次的」です：**
  $Var(aX + bY + c) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)$
- **共分散は線形です：** $Cov(aX + c, bY + d) = abCov(X, Y)$
  および $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$

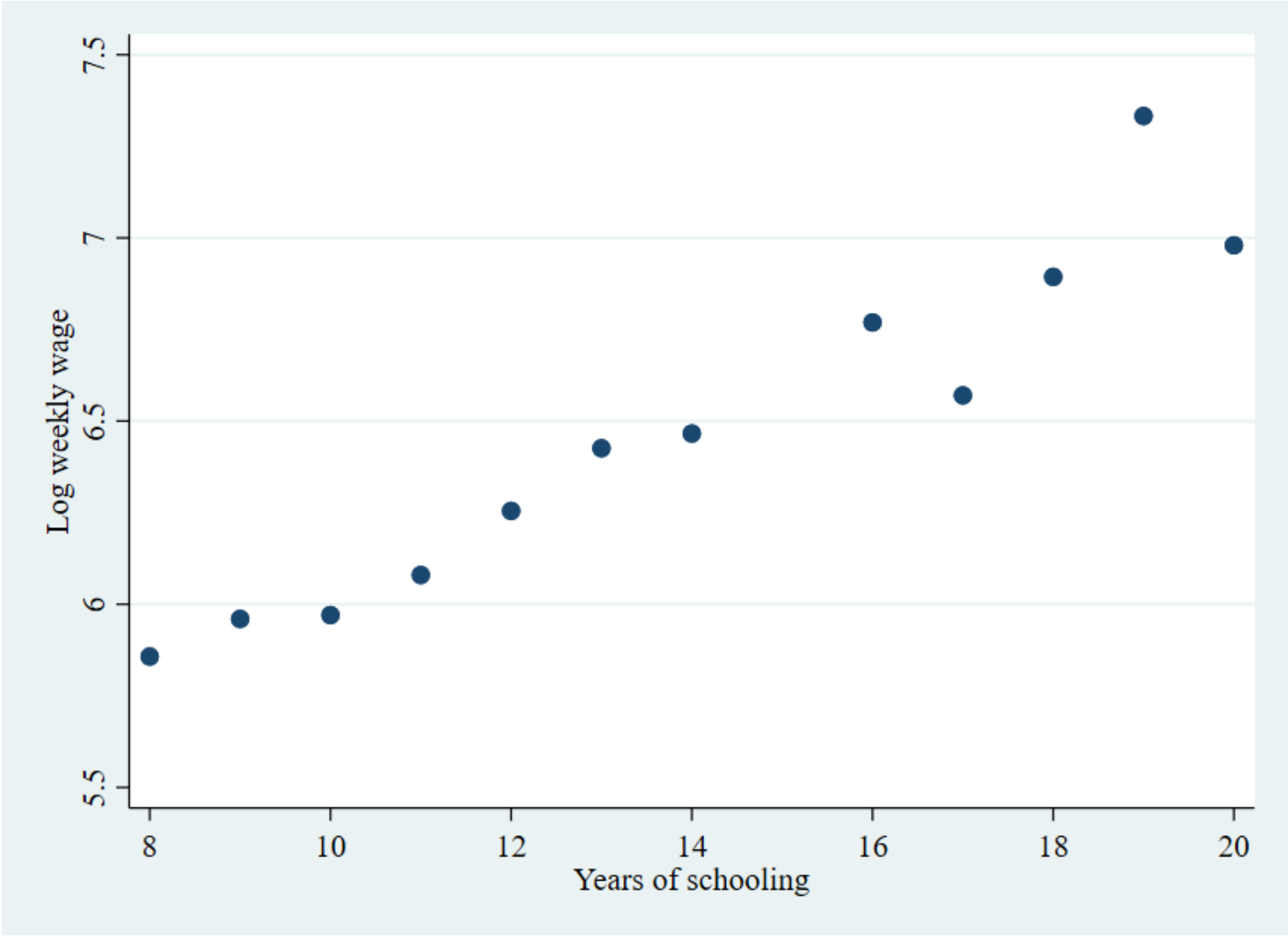
条件付き期待値

経済学では、特に条件付き期待値（conditional expectations）に興味があります。

- $X = x$ のときの Y の平均は何ですか？（例：ブラウン大学に進学した人の平均所得は？）
- 条件付き期待値関数 (CEF)：
  $E[Y \mid X = x] = \sum_{y \in \mathbb{Y}} y p(y \mid x)$（離散型の場合）
  $E[Y \mid X = x] = \int_{y \in \mathbb{Y}} y f(y \mid x) dy$（連続型の場合）
- X で評価されたランダムな CEF を $E[Y \mid X]$ と書くことがあります。
- すべての x に対して $E[Y \mid X = x] = E[Y]$ であるとき、Y は X について平均独立（mean independent）であると言います。
- X で条件付けると、X の関数は定数として扱えます：例：
  $E[f(X) + g(X)Y|X = x] = f(x) + g(x)E[Y|X = x]$。

条件付き期待値の例

教育年数を条件とした年収（対数）の期待値関数 (CEF)



反復期待値の法則 (LIE)

非常に重要な結果：反復期待値の法則（Law of Iterated Expectations, LIE）

例から始めましょう。米国の成人の平均身長を計算したいとします。LIEによれば、以下の手順で計算できます：

1) 男性の平均身長を計算する。

2) 女性の平均身長を計算する。

3) 男女それぞれの平均身長を（男女比で重み付けして）平均する。

数学的には：

$$
\begin{aligned}
E[\text{身長}] &= P(\text{女})E[\text{身長}|\text{女}] + P(\text{男})E[\text{身長}|\text{男}] \\
&= E[E[\text{身長}|\text{性別}]]
\end{aligned}
$$

反復期待値の法則 (LIE)

反復期待値の法則（**LIE**）の形式的な表現は：

$$E[Y] = E[E[Y \mid X]]$$

左辺の期待値は $p(y)$ を用い、右辺の外側の期待値は $p(x)$ を、内側の期待値は $p(y \mid x)$ を用いることに注意してください。

平均独立と無相関

LIE を用いると、平均独立が無相関を意味することを示せます：

$$\begin{aligned} Corr(X,Y) &\propto E[(X-E[X])(Y-E[Y])] \\ &= E[E[(X-E[X])(Y-E[Y]) \mid X]] \\ &= E[(X-E[X])E[Y-E[Y] \mid X]] \\ &= E[(X-E[X])(E[Y \mid X]-E[Y])] \\ &= 0 \text{（}E[Y \mid X]=E[Y] \text{ のとき）} \end{aligned}$$

各ステップがどのように導出されたか確認してください。

逆は成り立ちません：無相関な変数が平均従属であることもあります。

- もちろん、独立 $\Longrightarrow$ 平均独立 です（が、逆は必ずしも成り立ちません）。

ベクトル・行列表記についての補足

確率ベクトル $\mathbf{X} = [X_1, \ldots, X_N]'$ を扱うと便利なことが多いです。

- このコースでは、これらは常に「列ベクトル」とし、太字で表記します。
- 行列も太字で表記します（K 列 N 行など）。

ベクトル・行列の標準的な演算（転置、逆行列など）については、The Matrix Cookbook が参考になります。

- `www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

期待値は各要素ごとに適用されます：例：

$$E\left[\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}\right] = \begin{bmatrix} E[X_{11}] & E[X_{12}] \\ E[X_{21}] & E[X_{22}] \end{bmatrix}$$

- 分散は $Var\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{bmatrix}$、共分散は $Cov\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}\right) = \begin{bmatrix} Cov(X_1, Y_1) & Cov(X_1, Y_2) \\ Cov(X_2, Y_1) & Cov(X_2, Y_2) \end{bmatrix}$ のように定義されます。

アウトライン

1. 確率変数と確率分布 ✓
2. 平均と分散 ✓
3. **実験における識別**
4. 無作為抽出と標本平均
5. 仮説検定と推論

実験における識別

- これまでに学んだ理論を使えば、なぜ実験が（少なくとも識別の観点から）「うまくいく」のかを数学的に示すことができます。

- 潜在的結果（**potential outcomes**）の枠組みを思い出しましょう：
  - $Y_i(1), Y_i(0)$ は、個人 i が処置を受けた場合と受けなかった場合の結果です。
  - これらを用いて観測される結果をモデル化します：
    $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

- 平均処置効果（ATE）に興味があるとします：

$$ATE = E[Y_i(1) - Y_i(0)]$$

- 各個人にコイン投げで D_i を割り当てるとします。これは $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ を意味します。なぜでしょうか？

ランダム化の利用

- 実験により、$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ が成り立っています。
- $\underbrace{E[Y_i|D_i = 1]}_{\text{処置群の母平均}}$ はいくらでしょうか？

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1] = E[Y_i(1)]$$

最初の等号は潜在的結果モデルに基づき、2 番目の等号は（平均）独立性に基づきます。

- 同様に、$E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0] = E[Y_i(0)]$ となります。
- これらの結果を組み合わせると：

$$\underbrace{E[Y_i|D_i = 1]}_{\text{処置群の母平均}} - \underbrace{E[Y_i|D_i = 0]}_{\text{対照群の母平均}} = \underbrace{E[Y_i(1) - Y_i(0)]}_{\text{平均処置効果}} = \tau$$

したがって、実験における処置群と対照群の母平均の差によって ATE が識別されます！

条件付き非交絡性（Unconfoundedness）の下での識別

- ここで、$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|\mathbf{X}_i$ であると仮定します。ただし $\mathbf{X}_i$ は観察可能な特徴のベクトルです。

- これは条件付き非交絡性、あるいは観測される変数に基づく選択（**selection on observables**）と呼ばれます。

- 直感的には、条件付き非交絡性とは、同じ $\mathbf{X}_i$ の値を持つ人々の間では、実質的に実験が行われているのと同じ状況であることを意味します。
  - これは（無条件の）独立性から導かれますが、より弱い条件です。$\mathbf{X}_i$ を通じた非ランダム性を許容するからです。

- なぜ条件付き非交絡性を信じられるのでしょうか？
  - 層別実験（Stratified experiment）：同じ $\mathbf{X}_i$ を持つ人々の間でランダム化を行う場合。
  - 「擬似実験（Quasi-experiment）」/「自然実験」：同じ $\mathbf{X}_i$ を持つ人々の間では、D_i が実質的にランダムに決まっていると考えられる場合。

例：猛暑日とテストの点数

- Park et al (2021) は、学期中の猛暑日（D_i）がテストの点数（Y_i）に与える影響を研究しました。
  - 彼らの D_i は 2 値変数ではありませんが、例えば $D_i = 1[\text{猛暑日} > 10\text{日}]$ のように 2 値化して考えることもできます。
- なぜ $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$ だと考えられるのでしょうか？
  - 気候の異なる場所に住む人々は、他の側面（経済状況など）でも異なっている可能性があるからです。
- Park らは、X_i をその場所の過去の気象データとしたとき、$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|X_i$ であると主張しています。
  - 過去の気象パターンを条件とすれば、特定の年の暑さは実質的にランダムであるという考えです。
- これは妥当な仮定だと思いますか？

Park et al. (2021) 要約

Human capital generally, and cognitive skills specifically, play a crucial role in determining economic mobility and macroeconomic growth. While elevated temperatures have been shown to impair short-run cognitive performance, much less is known about whether heat exposure affects the rate of skill formation. We combine standardized achievement data for 58 countries and 12,000 US school districts with detailed weather and academic calendar information to show that the rate of learning decreases with an increase in the number of hot school days. These results provide evidence that climatic differences may contribute to differences in educational achievement both across countries and within countries by socioeconomic status and that may have important implications for the magnitude and functional form of climate damages in coupled human–natural systems.

例：大学の選択度（Selectivity）の収益

- Dale and Krueger (2002) は、私たちの例と似た問いを研究しました：選択度の高い大学に通うことは、所得にどのような影響を与えるか？

- 明らかに $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$ です。なぜなら、選択度の高い大学に通う学生は、もともとの学力も高い傾向があるからです。

- Dale と Krueger は、$\mathbf{X}_i$ を「その学生が出願し、合格した大学の集合」としたとき、$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|\mathbf{X}_i$ であると主張しました。

- 本質的に、彼らは「どの大学に出願し、どこに合格したか」さえ分かれば、その中からどこに進学するかは実質的にランダムであると議論しています。

- これを信じますか？ この仮定が崩れるとしたら、どのような理由が考えられますか？
  - 合格した大学の中からどこを選ぶかは、家庭環境、意欲、将来の計画などの違いを反映しているかもしれません。

Dale and Krueger (2002)

Abstract

Estimates of the effect of college selectivity on earnings may be biased because elite colleges admit students, in part, based on characteristics that are related to future earnings. We matched students who applied to, and were accepted by, similar colleges to try to eliminate this bias. Using the College and Beyond data set and National Longitudinal Survey of the High School Class of 1972, we find that students who attended more selective colleges earned about the same as students of seemingly comparable ability who attended less selective schools. Children from low-income families, however, earned more if they attended selective colleges.

条件付き非交絡性の利用

- $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$ であるとします。
- 実験の場合と同様に、すべての $\mathbf{x}$ に対して次が成り立ちます：

$$E[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}] = E[Y_i(1) | \mathbf{X}_i = \mathbf{x}]$$

および

$$E[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}] = E[Y_i(0) | \mathbf{X}_i = \mathbf{x}]$$

- これは次を意味します：

$$\underbrace{E[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\mathbf{X}_i = \mathbf{x} \text{ の処置群平均}} - \underbrace{E[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}]}_{\mathbf{X}_i = \mathbf{x} \text{ の対照群平均}} = \underbrace{E[Y_i(1) - Y_i(0) | \mathbf{X}_i = \mathbf{x}]}_{\mathbf{X}_i = \mathbf{x} \text{ での ATE}}$$

- $E[Y_i(1) - Y_i(0) | \mathbf{X}_i = \mathbf{x}]$ はしばしば条件付き平均処置効果（**CATE**）と呼ばれ、$CATE(\mathbf{x})$ と書かれます。

条件付き非交絡性の利用（続き）

- 条件付き非交絡性の下で、$CATE(\mathbf{x})$ が識別されることを示しました。
- では、無条件の $ATE = E[Y_i(1) - Y_i(0)]$ も識別されるでしょうか？
- はい！ 反復期待値の法則（LIE）より：

$$E[\underbrace{E[Y_i(1) - Y_i(0)|\mathbf{X}_i]}_{CATE(\mathbf{X}_i)}] = E[Y_i(1) - Y_i(0)]$$

- 技術的な注意：ここでは、すべての $\mathbf{x}$ について条件付き期待値が存在すると仮定しています。
- これには、$0 < Pr(D_i = 1|\mathbf{X}_i = \mathbf{x}]) < 1$ という条件が必要です。これはオーバーラップ（**overlap**）条件と呼ばれます。
- 直感的には、全体の ATE を知るためには、すべての X_i の値において処置群と対照群の両方が存在する必要があります。

母平均について学ぶ

- 実験において、平均処置効果が母平均の差として識別されることを示しました：

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1) - Y_i(0)]$$

- 同様に、条件付き非交絡性の下では、CATE は条件付き母平均の差によって識別されます。

- しかし、現実には母集団全体のデータは見えないため、$E[Y_i|D_i = 1]$ や $E[Y_i|D_i = 0]$ などを直接知ることはできません。

- 私たちは、観察されたサンプルからこれらの推定対象（**estimands**）について学ぶ必要があります。

- ここから統計的推論の出番です...

アウトライン

1. 確率変数と確率分布 ✓
2. 平均と分散 ✓
3. 実験における識別 ✓
4. **無作為抽出と標本平均**
5. 仮説検定と推論

サンプルの定義

- 統計的推論を定式化するために、観察されたデータが母集団からどのように抽出されたかを特定する必要があります。

- 基本的なケース：サイズ N の独立で同一な分布（**iid**）に従う代表的な（**representative**）サンプルを観察するとします：例：
  $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
  - 独立（**Independent**）：すべての $i \neq j$ について Y_i は Y_j と独立。
  - 同一な分布（**Identically distributed**）：すべての i, j について Y_i と Y_j は同じ分布に従う。
  - 代表的（**Representative**）：Y_i の分布が、関心のある母集団の分布と同じである。

- iid かつ代表的なデータは分析が容易な基準となりますが、現実には必ずしも成り立たないことを意識する必要があります。
  - 同じ世帯の人々を一緒に抽出する場合、独立ではありません！
  - 州ごとに層化抽出を行う場合、同一分布ではありません。
  - デューイ対トルーマンの例では、代表的ではありません！

標本平均の期待値と分散

母平均 $\mu = E[Y_i]$ を、サイズ N の iid かつ代表的なサンプル $\mathbf{Y}$ から学びたいとします。

- 自然な推定量は標本平均です：$\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ はランダムなデータ $\mathbf{Y}$ の関数です。したがって、それ自体が確率変数であり、分布（「標本分布」）を持ちます。

これまでに学んだことを使って、$\hat{\mu}$ の期待値と分散を導出できます：

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_i Y_i\right) = \frac{1}{N^2}\sum_i Var(Y_i) = \sigma^2/N$$

（ただし $\sigma^2 = Var(Y_i)$）

- 式 (1) は $\hat{\mu}$ が不偏であることを示しています：平均値は μ です。
- 式 (2) は $\hat{\mu}$ の分散がサンプルサイズ N とともに減少することを示しています（一致性に近い概念です）。

無作為抽出と標本平均

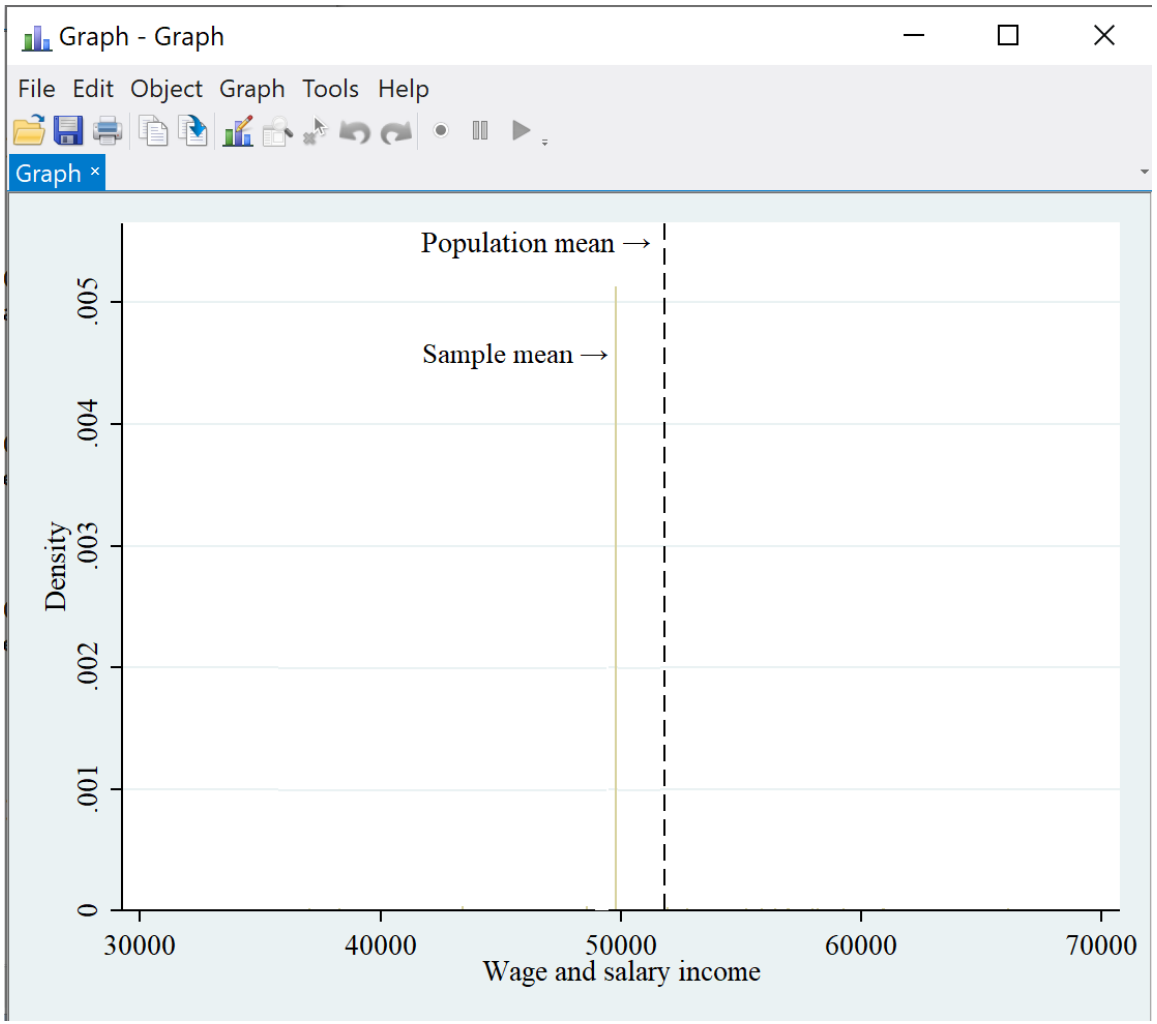
平均所得の推定における不偏性と一致性のシミュレーション：

Do-file Editor - Lecture2

File Edit View Project Tools

Lecture2 ×

```
1  summ incwage
2  local incwage_mean=r(mean)
3  matrix samp_means=J(50,1,.)
4  forval i=1/50 {
5      preserve
6      bsample 100
7      qui summ incwage
8      matrix samp_means[`i',1]=r(mean)
9      restore
10 }
11 preserve
12 clear
13 svmat samp_means
14 hist samp_means1, xline(`incwage_mean')
```



無作為抽出のシミュレーション

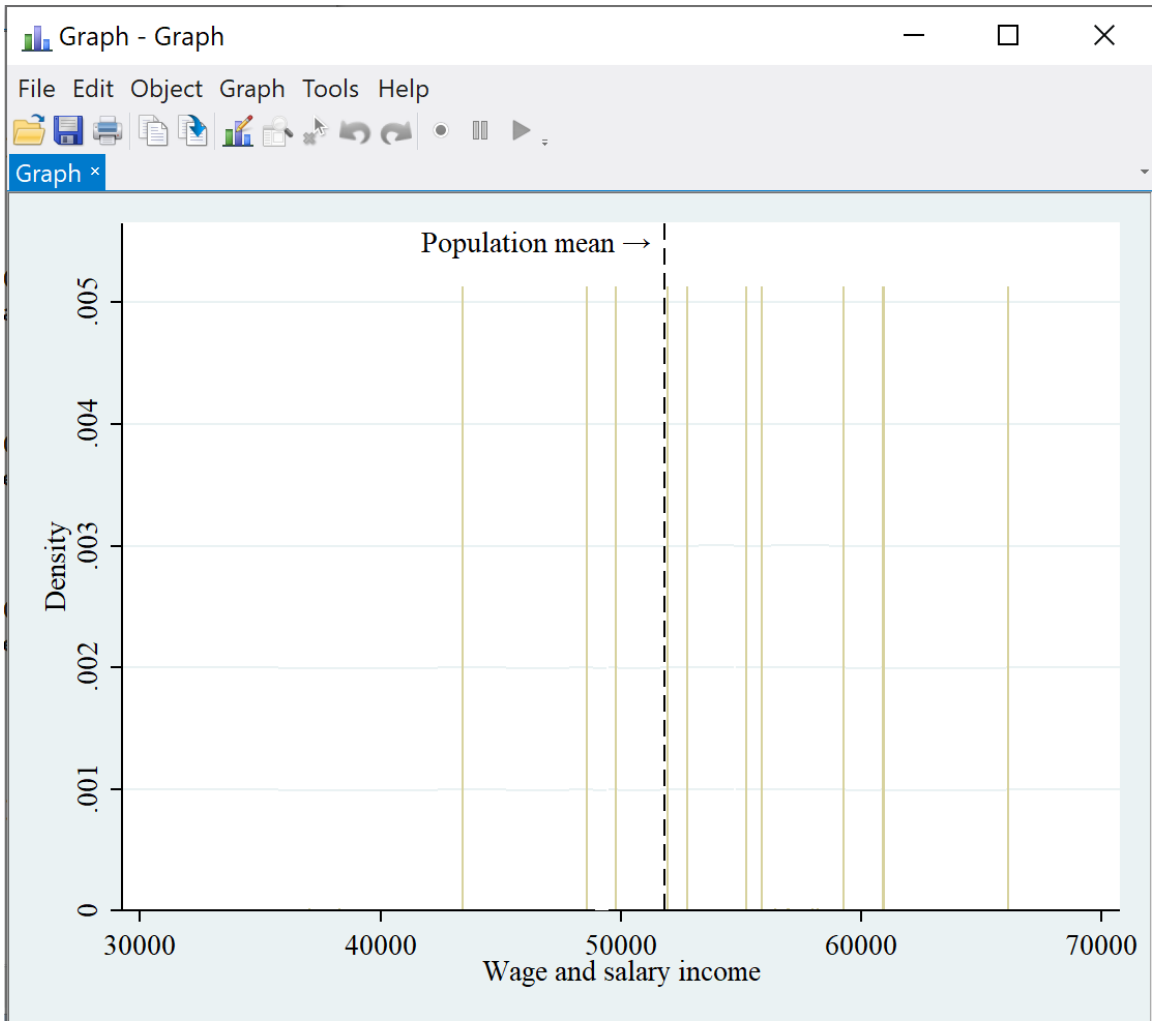
平均所得の推定における不偏性と一致性のシミュレーション：

Do-file Editor - Lecture2

File Edit View Project Tools

Lecture2 ×

```
 1  summ incwage
 2  local incwage_mean=r(mean)
 3  matrix samp_means=J(50,1,.)
 4  forval i=1/50 {
 5      preserve
 6      bsample 100
 7      qui summ incwage
 8      matrix samp_means[`i',1]=r(mean)
 9      restore
10  }
11  preserve
12  clear
13  svmat samp_means
14  hist samp_means1, xline(`incwage_mean')
```



無作為抽出と標本平均

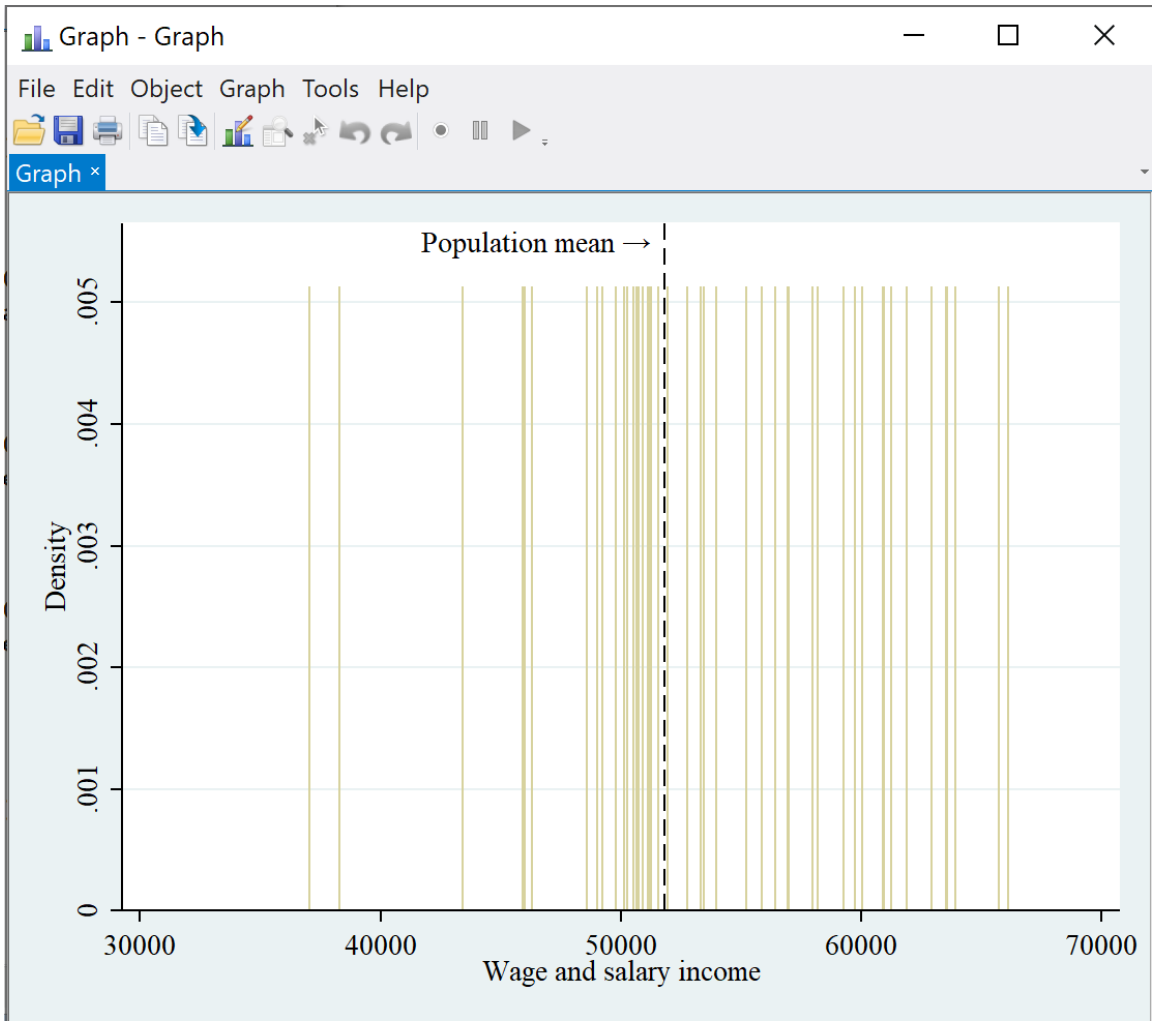
平均所得の推定における不偏性と一致性のシミュレーション：

Do-file Editor - Lecture2

File Edit View Project Tools

Lecture2 ×

```
 1  summ incwage
 2  local incwage_mean=r(mean)
 3  matrix samp_means=J(50,1,.)
 4  forval i=1/50 {
 5      preserve
 6      bsample 100
 7      qui summ incwage
 8      matrix samp_means[`i',1]=r(mean)
 9      restore
10  }
11  preserve
12  clear
13  svmat samp_means
14  hist samp_means1, xline(`incwage_mean')
```



無作為抽出と標本平均

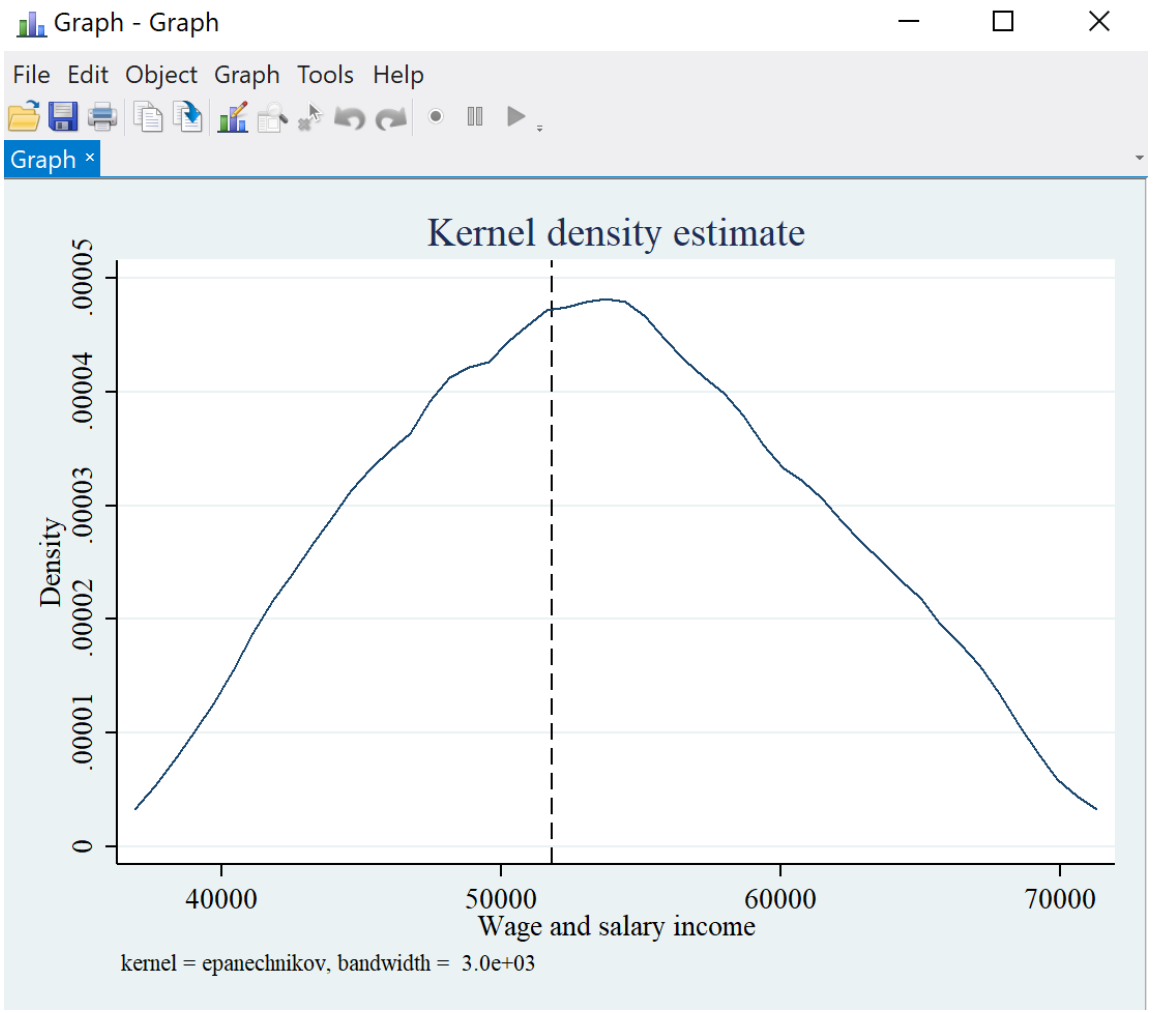
平均所得の推定における不偏性と一致性のシミュレーション：

Do-file Editor - Lecture2

File Edit View Project Tools

Lecture2 ×

```
 1  summ incwage
 2  local incwage_mean=r(mean)
 3  matrix samp_means=J(50,1,.)
 4  forval i=1/50 {
 5      preserve
 6      bsample 100
 7      qui summ incwage
 8      matrix samp_means[`i',1]=r(mean)
 9      restore
10  }
11  preserve
12  clear
13  svmat samp_means
14  hist samp_means1, xline(`incwage_mean')
```



無作為抽出と標本平均

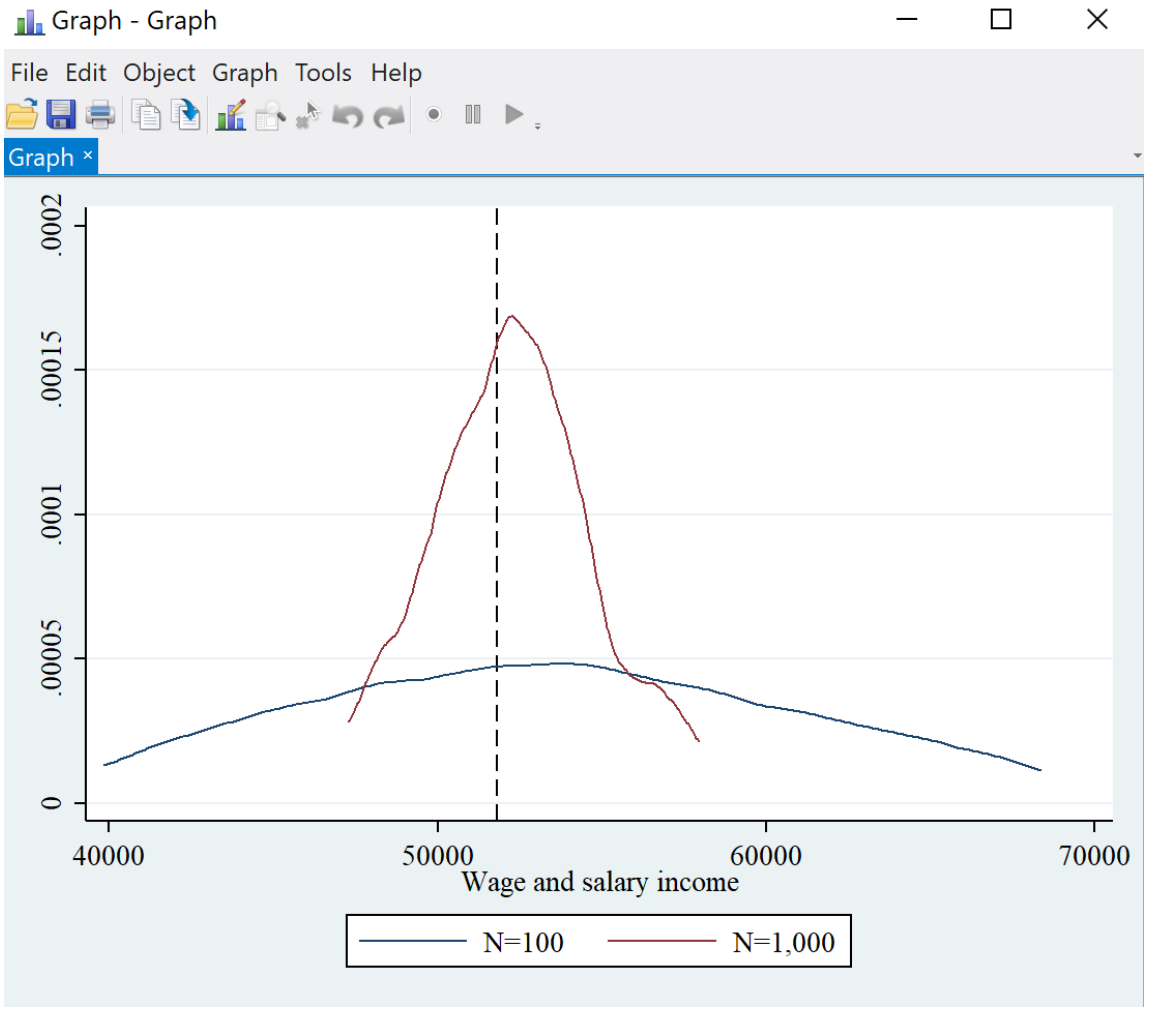
平均所得の推定における不偏性と一致性のシミュレーション：

Do-file Editor - Lecture2

File Edit View Project Tools

Lecture2 ×

```
1  summ incwage
2  local incwage_mean=r(mean)
3  matrix samp_means=J(50,1,.)
4  forval i=1/50 {
5      preserve
6      bsample 1000
7      qui summ incwage
8      matrix samp_means[`i',1]=r(mean)
9      restore
10 }
11 preserve
12 clear
13 svmat samp_means
14 hist samp_means1, xline(`incwage_mean')
```



無作為抽出と標本平均

まとめ：$\hat{\mu} = \frac{1}{N}\sum_i Y_i$ が $\mu = E[Y_i]$ の良い推定量である 2 つの理由：

- 不偏である：$E[\hat{\mu}] = \mu$
- サンプルサイズが大きくなるにつれて分散がゼロに収束する：$\lim_{N\to\infty} Var(\hat{\mu}) = 0$

次章では、$\hat{\mu}$ のもう一つの優れた性質を見ます：N が大きいとき、その分布は近似的に正規分布に従います。

無作為抽出と標本平均

条件付き期待値 $\mu(x) = E[Y_i \mid X_i = x]$ に関心がある場合、$X_i = x$ を条件とした Y_i の条件付き標本平均も考えられます。

- これは X_i が少ない値 x を持つ離散型の場合に最も簡単です。
- 自然な推定量 $\hat{\mu}(x) = \frac{1}{N_x} \sum_{i:X_i=x} Y_i$（ただし N_x は $X_i = x$ である観測数）。

先ほどと同様の導出により：

$$E[\hat{\mu}(x)] = E[Y_i \mid X_i = x] = \mu(x)$$
$$Var(\hat{\mu}(x)) = Var(Y_i \mid X_i = x)/N_x$$

したがって、これも不偏推定量であり、$N_x \to \infty$ となれば真の値に近づきます。

- X_i が離散型でない場合や、N_x が大きくならない場合はどうなるでしょうか？ それは後ほど...

アウトライン

1. 確率変数と確率分布 ✓
2. 平均と分散 ✓
3. 実験における識別 ✓
4. 無作為抽出と標本平均 ✓
5. 仮説検定と推論

仮説検定 – 導入

- N が大きくなれば、標本平均 $\hat{\mu}$ は母平均 μ に近づくことを示しました。
- しかし、「近い」とはどういう意味でしょうか？
- データの所得の標本平均が 50,000 ドルだったとき、母平均が 55,000 ドルであると考えるのは妥当でしょうか？ 70,000 ドルならどうでしょうか？
- 仮説検定（**Hypothesis testing**）は、この「近い」という概念を定式化するのに役立ちます。
- それは、もし真の値が 55,000 ドルや 70,000 ドルだった場合に、標本平均が 50,000 ドルになる確率がどの程度あるかを教えてくれます。

仮説検定の概要

1. 母平均が特定の値であるという帰無仮説を指定します：$H_0 : \mu = \mu_0$。
   - 例：母平均が 55,000 ドルなら $H_0 : \mu = 55,000$。
2. 帰無仮説が正しいとした場合に、$\hat{\mu}$ が μ_0 から少なくともこれほど離れた値として観察される確率を計算します。これを **p 値（p-value）**と呼びます。
3. p 値が小さい場合、帰無仮説を棄却（**reject**）します。すなわち、帰無仮説が正しいとしたら、これほど μ_0 から離れた $\hat{\mu}$ が観察されるのは考えにくいということです。
   - 一般的なしきい値は $\alpha = 0.05$ です。
4. このようにして棄却できないすべての μ_0 の値を集めて、信頼区間（**confidence interval, CI**）を形成します。
   - $\alpha = 0.05$ の場合、構築された信頼区間は、データの実現値の 95% において真の値 μ を含みます。

標本平均 $\hat{\mu}$ が正規分布に従う場合の仮説検定

- 特別なケースとして、$\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ かつ σ^2 が既知である場合を考えましょう。
- なぜこのケースを考えるのでしょうか？
  - $Y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$ なら、これが $\hat{\mu}$ の正確な分布になります。
  - 次章で示しますが、Y_i が正規分布でなくても、N が大きければ $\hat{\mu}$ は近似的に正規分布に従います。
  - また、N が大きければ σ^2 を非常に精度よく推定できることも示します。
- 帰無仮説 $H_0 : \mu = \mu_0$（例：平均所得は 55,000 ドル）を検定したいとします。
- $\hat{t} = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}}$ とすると、帰無仮説の下で $\hat{t} \sim \mathrm{N}(0,1)$ となります。
  - $\hat{t}$ の分布は、サンプル $(Y_1, ..., Y_N)$ の抽出を何度も繰り返したときの分布です。

標本平均 $\hat{\mu}$ が正規分布に従う場合の仮説検定（続き）

- 帰無仮説 $H_0 : \mu = \mu_0$ の下で、$\hat{t} \sim \mathrm{N}(0,1)$ であることを示しました。
- ある $t \geq 0$ に対して、$Pr(|\hat{t}| > t)$ はいくらでしょうか？

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

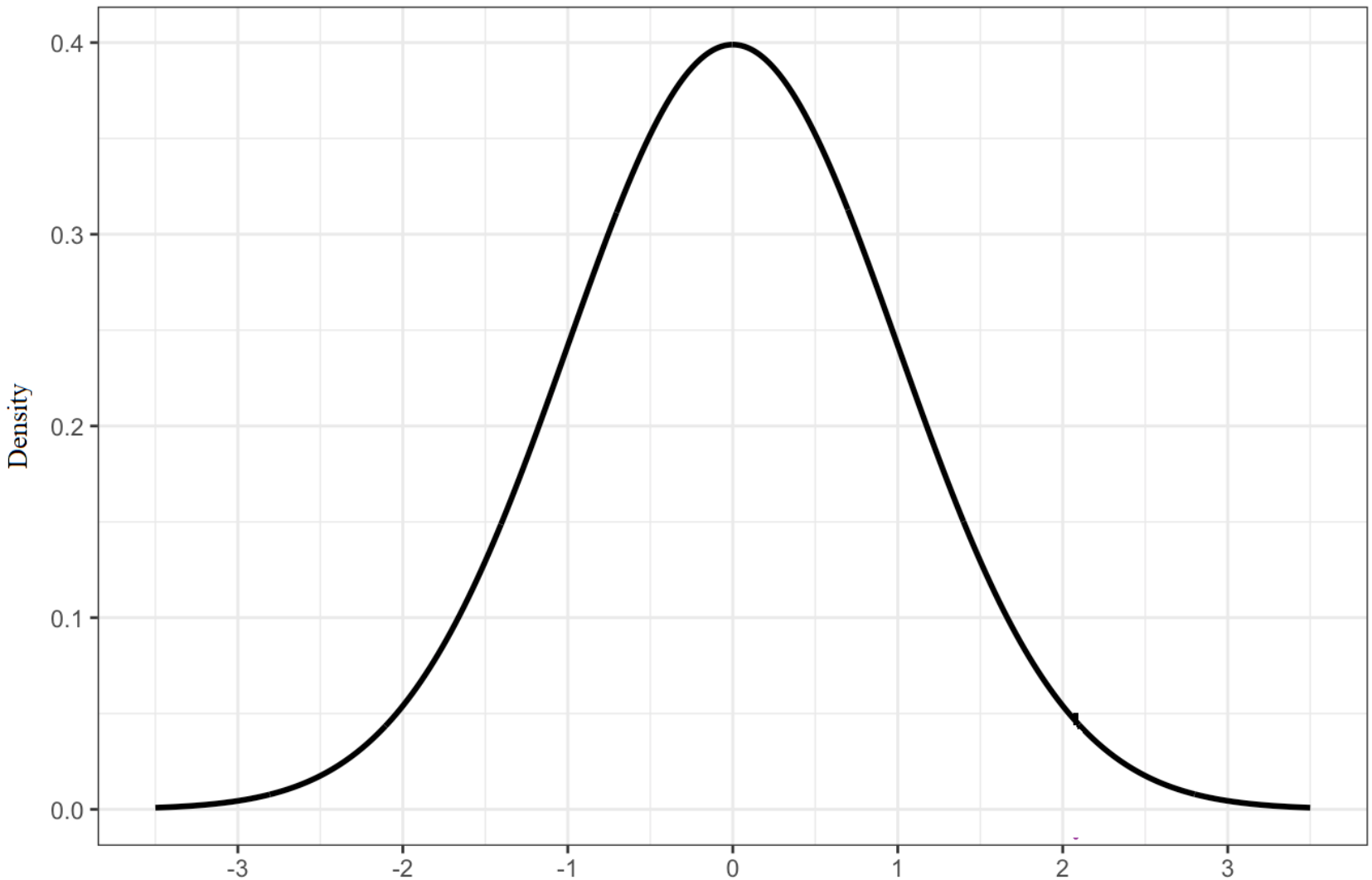
- 帰無仮説 $H_0 : \mu = \mu_0$ に対する p 値を次のように定義します：

$$\begin{aligned} p(\hat{t}) = 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)) &= 1 - \left(\Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \right) - \Phi\left(\frac{-|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \right) \right) \\ &= 2\left(1 - \Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \right) \right) \end{aligned}$$

- 直感的には、p 値は、帰無仮説が正しいときに少なくともこれほど大きな $|\hat{t}|$ が観察される確率です。

p 値構築のイラスト

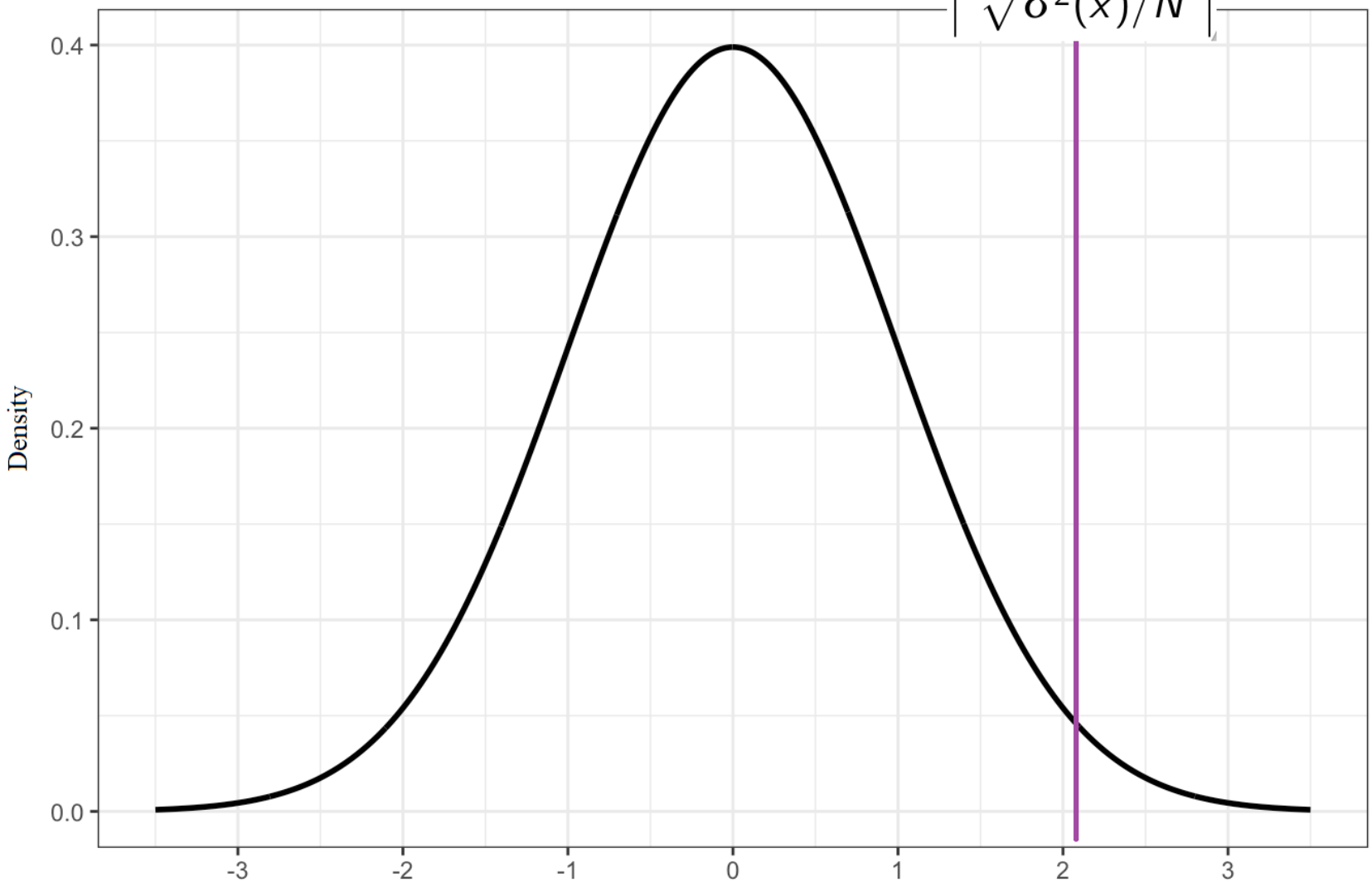
標準正規分布の密度関数（平均 0、標準偏差 1）



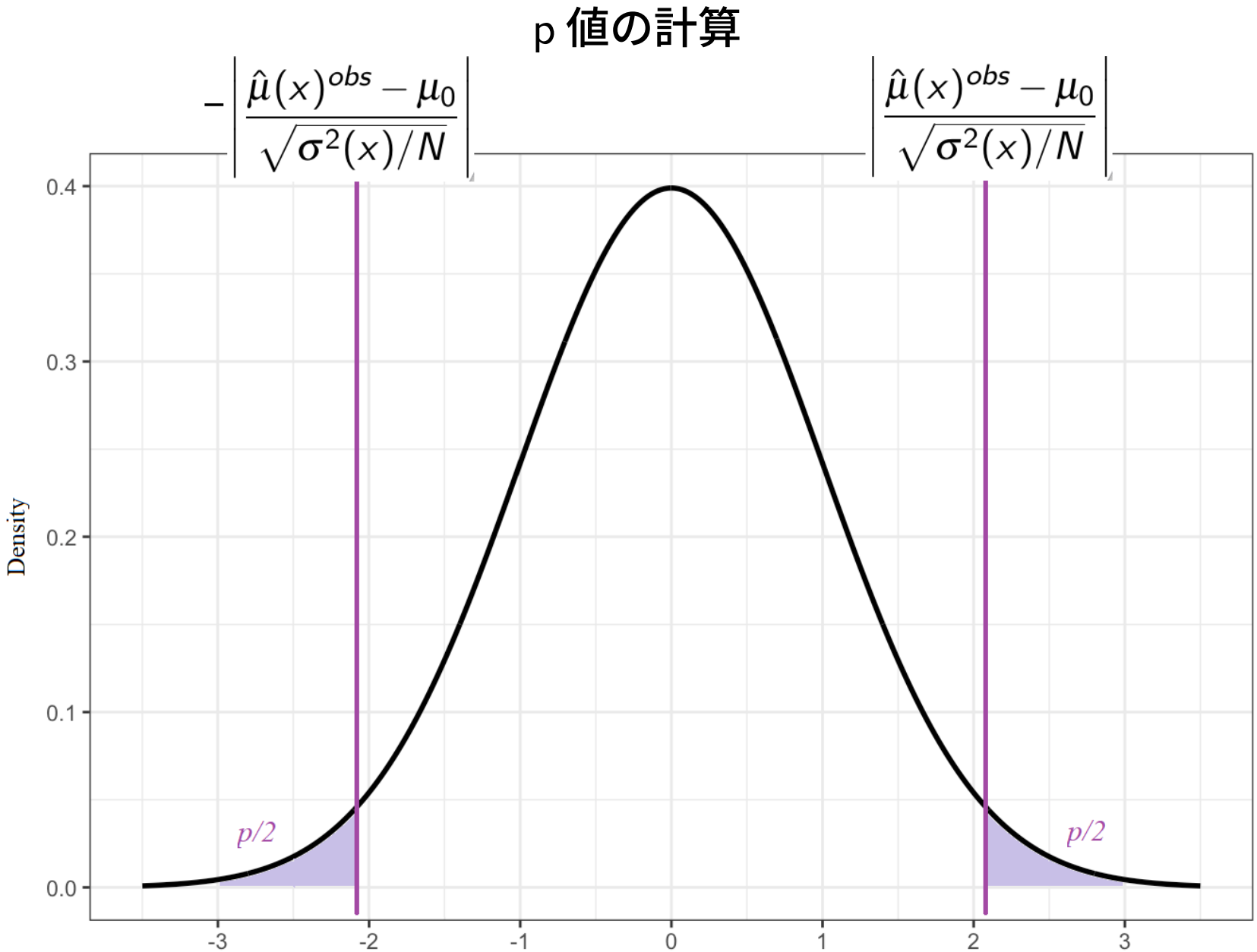
p 値構築のイラスト

ランダムな推定量を標準化した値（実現値）

$$\left| \frac{\hat{\mu}(x)^{obs} - \mu_0}{\sqrt{\sigma^2(x)/N}} \right|$$



p 値構築のイラスト


p 値の計算
$-\left|\frac{\hat{\mu}(x)^{obs}-\mu_0}{\sqrt{\sigma^2(x)/N}}\right|$
$\left|\frac{\hat{\mu}(x)^{obs}-\mu_0}{\sqrt{\sigma^2(x)/N}}\right|$
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
p/2
p/2

いつ帰無仮説を棄却するか？

- p 値の形式を思い出してください：

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- $\Phi(1.96) - \Phi(-1.96) \approx 0.95$ となることが知られています。したがって、$p < 0.05$ となるのは $|\hat{t}| > 1.96$ のとき、かつそのときに限られます。つまり、5%有意水準で $|\hat{t}| > 1.96$ なら棄却します。

- これは、どのような μ_0 を棄却し、どのような μ_0 を棄却しないことを意味するでしょうか？

- 次の場合に棄却しません：

$$|\hat{t}| \leq 1.96 \Longrightarrow \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96 \Longrightarrow \mu_0 \in [\hat{\mu} - 1.96\sigma/\sqrt{N}, \hat{\mu} + 1.96\sigma/\sqrt{N}]$$

- したがって、区間 $\hat{\mu} \pm 1.96\sigma/\sqrt{N}$ が 95% 信頼区間 (CI) となります。
  - これは、$H_0 : \mu = \mu_0$ が正しいときに $Pr(\mu_0 \in CI) = 0.95$ となる性質を持ちます。

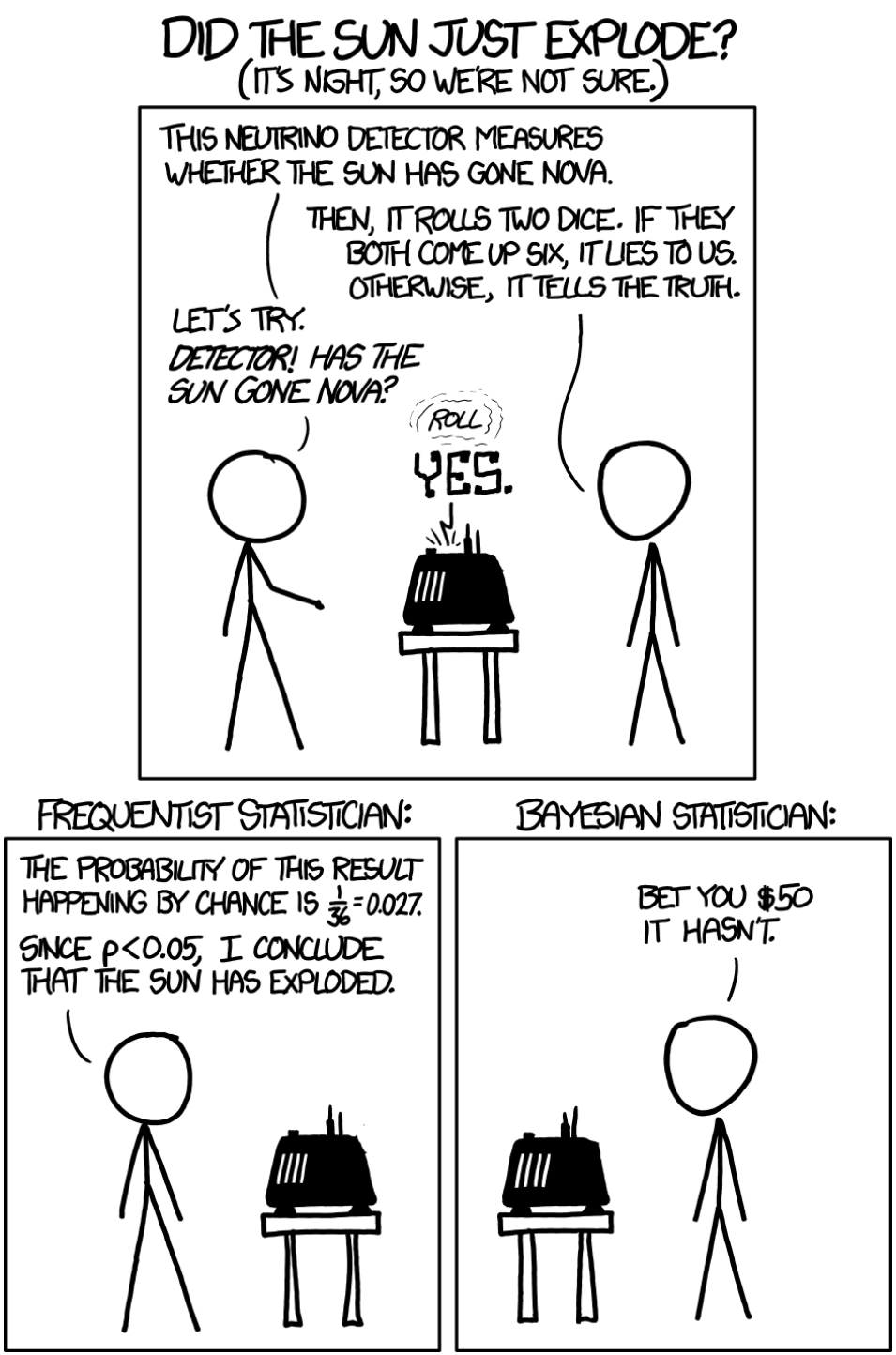
有意水準と検出力

- 検定の有意水準（**significance level**）（またはサイズ）とは、帰無仮説が正しいときに誤って棄却してしまうあらかじめ設定された確率です（第 I 種の過誤の発生率）。
  - 例：5%水準の検定は $p < 0.05$ のときに棄却します。

- 検定の検出力（**power**）とは、帰無仮説が偽であるときに正しく棄却できる確率です（1 - 第 II 種の過誤の発生率）。
  - 検出力は対立仮説の関数です。すなわち、実際には $\mu = \mu_A$ であるときに $H_0 : \mu = \mu_0$ を棄却する確率です。

p 値の解釈に関する注意

- 頻度論的な p 値はしばしば「帰無仮説が正しい確率」と解釈されますが、これは正しいでしょうか？ いいえ！
  - p 値は、帰無仮説が正しいと仮定した場合に、観察されたデータが得られる確率を示しています。
  - つまり、p 値は $P(data|H_0)$ について述べており、$P(H_0|data)$ ではありません。
  - ベイズの定理によれば、$P(H_0|Data) = P(Data|H_0) * P(H_0)/P(Data)$ です。これを定式化するには、帰無仮説が正しいという事前の信念 $P(H_0)$ を設定する必要があります。
- $p < 0.05$ なら「効果がある」強い証拠であり、$p \geq 0.05$ なら「効果がない」証拠であると解釈されがちです。
  - しかし、0.05 というしきい値はかなり恣意的なものです。
  - p 値は、帰無仮説の下で観察されたデータがどの程度尤もらしいかを示すスペクトラムとして捉えるのが適切です。
  - さらに、帰無仮説が偽であっても（検出力が低ければ）p 値が大きくなることもあります。


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › social selection › article

Published: 26 February 2015

Psychology journal bans *P* values

Chris Woolston

Nature **519**, 9 (2015) | Cite this article

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › social selection › article

Published: 26 February 2015

Psychology journal bans *P* values

Chris Woolston

Nature **519**, 9 (2015) | Cite this article

09 March 2015 This story originally asserted that "The closer to zero the P value gets, the greater the chance the null hypothesis is false." P values do not give the probability that a null hypothesis is false, they give the probability of obtaining data at least as extreme as those observed, if the null hypothesis was true. It is by convention that smaller P values are interpreted as stronger evidence that the null hypothesis is false. The text has been changed to reflect this.

非正規データの場合は？

- 分散が既知の正規確率変数の平均について、仮説検定や推論の方法が分かりました。... だから何だと言うのでしょうか？
  - ほとんどの変数は正規分布に従いません！
  - たとえ正規分布だとしても、なぜ分散が既知だと言えるのでしょうか？！
  - 私は皆さんの時間を無駄にさせてしまったのでしょうか！？　いいえ。
- 次に、強力な漸近（**asymptotic**）理論を概観します。これにより、サンプルが「十分に大きい」場合には、Y_i が正規分布に従わなくても、同様の推論ツールを適用できるようになります。