データ解析

第11回：統計的検定

渡辺澄夫

仮定したこと
は真実
だったのか

実世界

観測データ

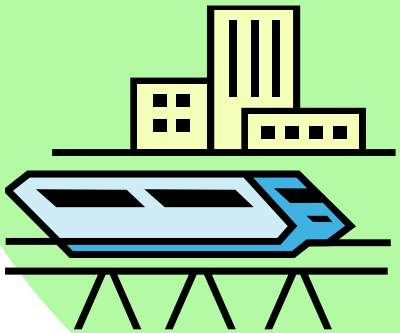
回帰・判別分析

推定と検定

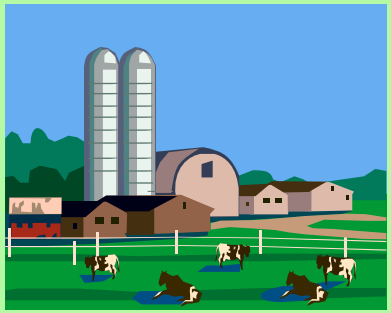
解析方法

統計的検定

ベイズ法
・階層ベイズ法



時系列予測



因子・主成分
・クラスタ分析

統計モデルは真ではない

データが得られたとき、データを発生した真の分布について考察するときにデータ解析や統計モデルが利用される。それらは数式で与えられるために、まるで正しい法則があたえられているかのように感じられるかもしれない。

しかしながら、それらはあくまでも人間が作った道具に過ぎない。

このことに気づいたとき、多くの人は次のように思うだろう。「自分たちが行なっている解析のもとである仮説が、そもそも正しくないのでは」。

どんなに多くても有限個のデータしかない状況で、この問いに確率1で答えることはできない。しかしながら、「可能な限り合理的にこの問いに答えることができるようにするにはどうしたらよいか」を取り扱うための基盤のひとつが統計的検定である。

統計的検定を学ぶためのコツ

実世界では、統計的検定はしばしば意思決定の基礎になるのであるが、
意思決定というものには、夢や希望や物語などが付随していることが多い。

もちろん検定は個人感情や企業利益を排して行なわれなければ意味がない。

このため統計的検定のしくみを説明する時には無味乾燥な用語が使われる。
それが学生のみなさんの「検定嫌い」の原因のひとつのような気がする。

統計的検定を始めて学ぶ時には、むしろ、意識的に物語を設定することで
「人間は物語だと理解しやすい」という特徴を利用しよう。そして統計的
検定の手順や用語を覚えよう。

いったん、物語が理解できれば、学生のみなさんは、その物語を客観的に
見ることができるようになる。その上で、統計的検定を実用する時には、
夢や希望や物語を排して、厳正に適用するようにしよう。

統計的検定が必要になる状況とは

典型的な物語

ラーメン屋さんを経営しているあなたは、より多くの人にラーメンを食べてもらいたいという夢を持っていた。そこで、スープ・麺・チャーシュー・メンマ卵などを求めて日本全国をめぐり最高の素材を見つけ出した。また素材の良さを最も活かすための調理法を求めて、人口調味料はいっさい使わない最高のダシの取り方を探求し雨の日も風の日も努力を続けた。

こうして何年もの努力の末、とうとうあなたは究極至高のラーメンを作り出すことに成功した。そのラーメンをメニューに加えて、多くの人が来店してくれることを待つことにした。

これまでは、あなたのお店に来る人は1日平均100人標準偏差が10人だった。究極至高のラーメンを出品した後５日間で来店した人は 110, 100,120, 90, 130だった。これまでよりも多くの人がお店に来てくれるようになったのだろうか。

検定の用語を覚えよう(1)。

仮説「お店に来る人はこれまでと変わらない」

＝データはいままでと同じ確率分布から出た

＝帰無仮説 (H0, Null Hypothesis)

仮説「お店に来る人はこれまでより多くなった」

＝データを発生した確率分布は変化した

＝対立仮説 (H1, Alternative Hypothesis)

統計的検定で何をしたいのか

ラーメン屋さんの例では究極至高のラーメンを出した後に来店した人数は
110, 100,120, 90, 130 だった。従来よりも多くなったように見える。

あなた:「来店者数が増えた」(対立仮説)と主張したい。

反論者:「帰無仮説のもとでデータは統計的ばらつきとして無理なく説明可能」。

あなた:「帰無仮説が正しいと仮定して、データが生じる確率が基準値よりも
小さいことを示します」

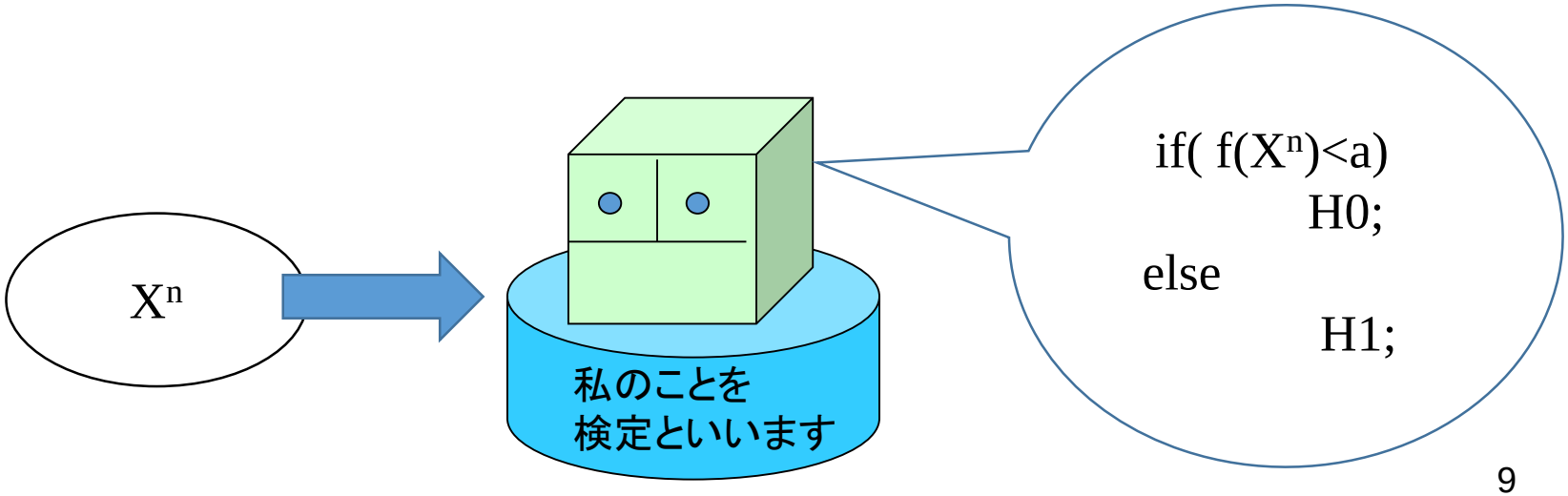
◎　背理法の基本構造:命題Pを示したいとき「 not P」を仮定して矛盾を
みちびけばよい。

◎　統計的検定の基本構造:対立仮説を示したいとき帰無仮説を仮定して
データがものすごく小さい確率でしか起こらないことを示す。

検定の用語を覚えよう(2)。

データを　$X^n = X_1, X_2, \ldots, X_n$　とする。

定義.「統計的検定」とは関数「$f : X^n \rightarrow$ 実数」と実数 a のペアを用いて
【もし $f(X^n)<a$ ならば H0 を結論とし、そうでないときは H1 を結論とする】
と定められたルールのことである。



（注）どんなルールでも、ひとつの検定である。

棄却域

関数 f: $X^n \rightarrow \{H0,H1\}$ と a のペアをひとつ固定して、検定を作った。

帰無仮説を捨てて対立仮説を取るという結果を与えるデータの
集合は $\{X^n ; f(X^n) \geqq a\}$ である。これを 棄却域 という。

（例）ラーメン屋さんの話では、５日間の来客数が X_1, X_2, X_3, X_4, X_5 のとき、例えば $Y = f(X^n) = X_1+X_2+X_3+X_4+X_5$ として a=550 とすると

if ($X_1+X_2+X_3+X_4+X_5<550$)
　　H0;
else
　　H1;

このとき棄却域は $\{X^n ; X_1+X_2+X_3+X_4+X_5 \geqq 550\}$ である。
$\{Y;Y \geqq 550\}$ と書いてもよい。

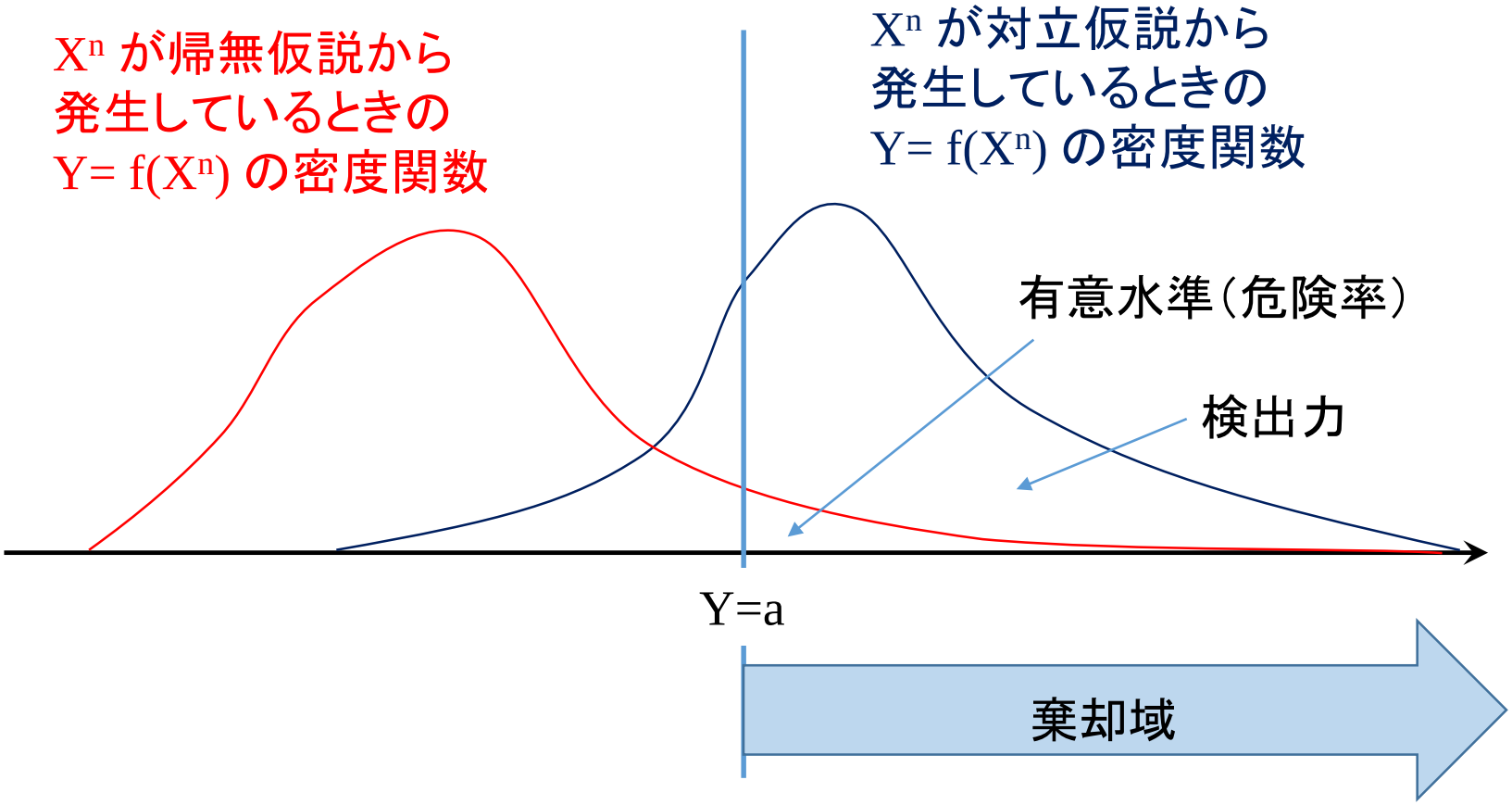
有意水準と検出力

(1)
帰無仮説が正しいとき対立仮説を選ぶことを第１種の誤りという。
また、その確率を危険率あるいは有意水準という。

(2)
対立仮説が正しいのに帰無仮説を選ぶことを第２種の誤りという。
対立仮説が正しいとき対立仮説を選ぶ確率を検出力(パワー)という。

（例）ラーメン屋さんの話では、真実は変わっていないのに
来客数が多くなったと判定する確率が危険率である。
一方、真実が変わったとき、確かに変化したことと
判定することが検出力である。

言葉の意味



通常は有意水準（危険率）を十分に小さく設定する。
しかし有意水準をあまりに小さくすると検出力も小さくなってしまい
検定を行なう意味がなくなってしまう。

「優れた検定」とは

関数「$f: X^n \rightarrow$ 実数」と実数 a のペアなら何でも検定であるが・・・

〇危険率が小さいほどよい検定である。

〇検出力が大きいほどよい検定である。

〇危険率も検出力も、同じ集合（棄却域）にデータが入る確率であるから危険率をいくらでも小さくし、かつ、検出力をいくらでも大きくすることはできない。

〇できるだけ危険率が小さく、かつ、できるだけ検出力が大きい検定を作るためには、関数 f をどのようにしたらよいのだろうか。

〇 関数 f として最も良いもの（最強検定、来週述べる）が存在する場合もあるが、一般には一番よいといえるものはないので関数の作り方には任意性が含まれている。

検定を作る例

帰無仮説：X^n は平均 m_0 分散 σ^2 の正規分布から発生した。
対立仮説：X^n は平均 m_1 $(m_1>m_0)$ 分散 σ^2 の正規分布から発生した。

(1) 関数 $f(X^n) = \Sigma (X_i - m_0) / (\sigma n^{1/2})$ を用いてみよう。

(2) 帰無仮説が正しいとき $Y=f(X^n)$ は平均 0 で分散が1の正規分布に従う。

(3) 対立仮説が正しいとき $Y=f(X^n)$ は平均 $n^{1/2}(m_1-m_0) / \sigma$ で分散が1の正規分布に従う。

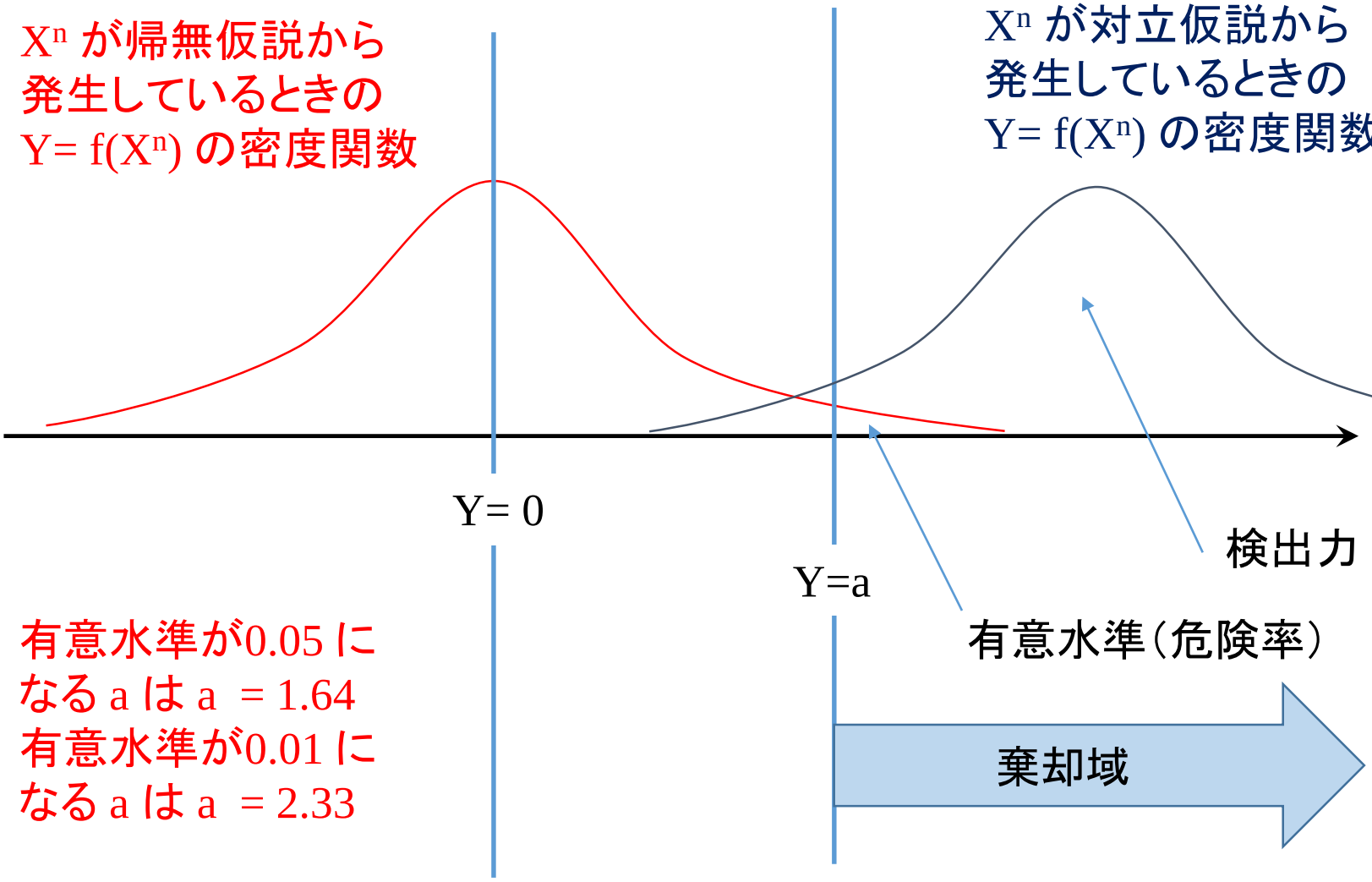
(4) 有意水準に対応させて棄却域を定める。$p(x)$が平均0分散1の正規分布のとき

$$0.05 = \int_{-\infty}^{a} p(x)\,dx \qquad \text{をみたす } a \text{ は } a=1.64$$

$$0.01 = \int_{-\infty}^{a} p(x)\,dx \qquad \text{をみたす } a \text{ は } a=2.33$$

片側検定

対立仮説が($m_1 > m_0$) のときは片側検定になる。



有意水準が0.05 に
なる a は a = 1.64
有意水準が0.01 に
なる a は a = 2.33

具体的な例

ラーメン屋さんの例に適用してみよう。

帰無仮説：X^n は平均 100 分散 10^2 の正規分布から発生した。

対立仮説：X^n は平均 m_1 ($m_1>m_0$) 分散 σ^2 の正規分布から発生した。

帰無仮説が正しいとき $Y=(X1+X2+X3+X4+X5-500) / (10*5^{1/2})$ は
　平均0分散1の正規分布に従う。棄却域は $Y>2.33$.

$Y=(110+100+120+90+130-500) / (10*5^{1/2}) =2.2$　従って

有意水準 0.05 で「多くのお客さんが来るようになった」
有意水準 0.01 で「多くのお客さんが来るようになったとはいえない」

と判定された。

両側検定

対立仮説が($m_1 \neq m_0$) のときは両側検定になる。

帰無仮説：X^n は平均 m_0 分散 σ^2 の正規分布から発生した。
対立仮説：X^n は平均 m_1 ($m_1 \neq m_0$) 分散 σ^2 の正規分布から発生した。

(1)　関数 $f(X^n) = \Sigma (X_i - m_0) / (\sigma n^{1/2})$ を使う（正確には $|f(X^n)|$ を使う）。

(2) 帰無仮説が正しいとき $Y=f(X^n)$ は平均 0 で分散が1の正規分布に従う。

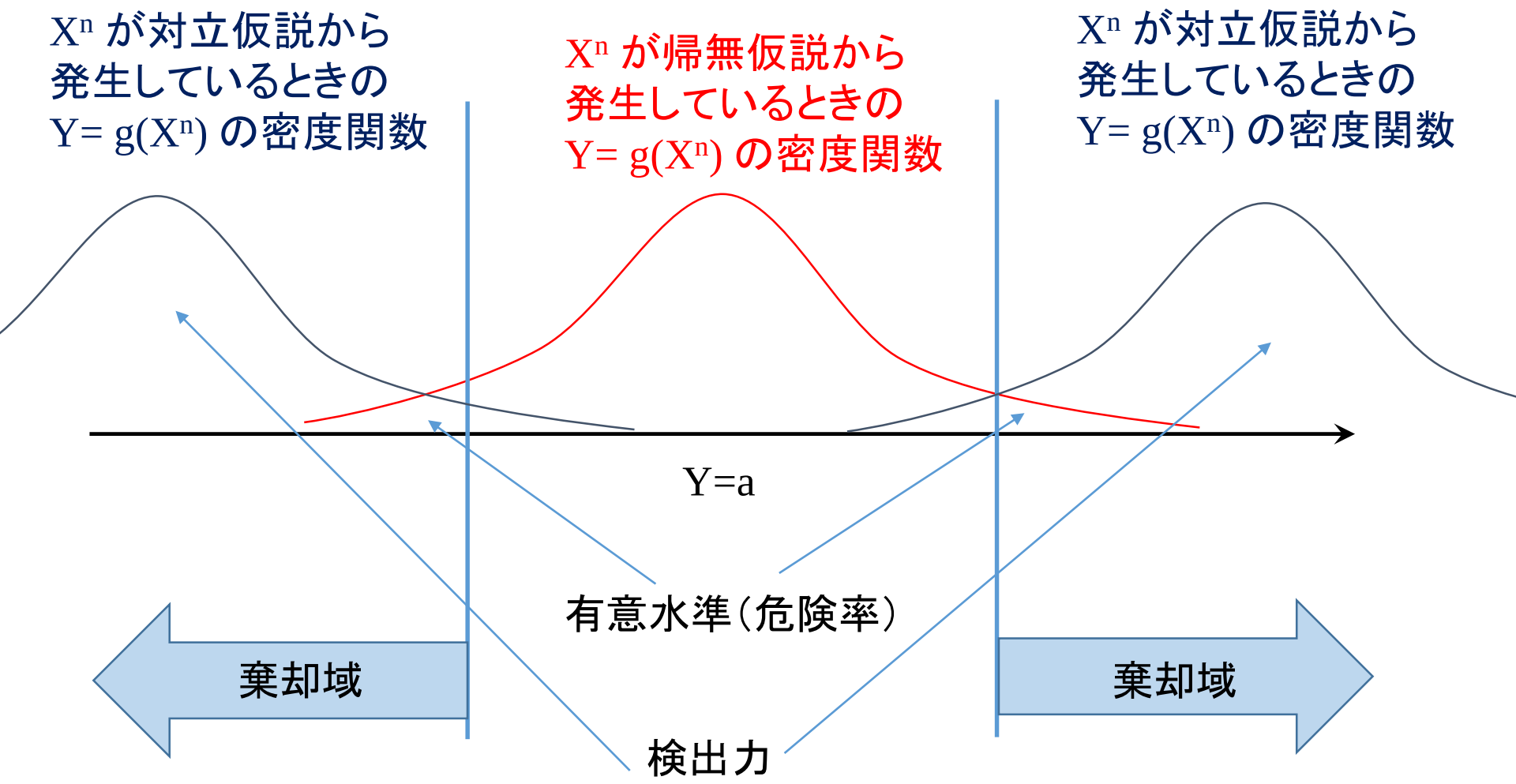
(3) 対立仮説が正しいとき $Y=f(X^n)$ は平均 $n^{1/2}(m_1-m_0) / \sigma$ で分散が1の
正規分布に従う。

(4) 有意水準に対応して棄却域を定める。

$$0.05 = \int_{-a}^{a} p(x)\,dx$$ をみたす a は a=1.96

$$0.01 = \int_{-a}^{a} p(x)\,dx$$ をみたす a は a=2.58

両側検定


X^n が対立仮説から
発生しているときの
$Y= g(X^n)$ の密度関数
X^n が帰無仮説から
発生しているときの
$Y= g(X^n)$ の密度関数
X^n が対立仮説から
発生しているときの
$Y= g(X^n)$ の密度関数
Y=a
有意水準（危険率）
棄却域
棄却域
検出力

関数 $f(X^n)$ と棄却域はどのように作るのか

「検定を作る手順」

(1)　関数 $Y=f(X^n)$ を定めて X^n が帰無仮説に従うときの Y の
密度関数を導出する。
(2) 有意水準と対立仮説から棄却域を定める。
(3) 実際のデータについて検定を行なう。

ここで、関数 $f(X^n)$ と棄却域の定め方には、任意性が含まれている。
二つの検定があるとき、優位水準が小さくて検出力が大きくなるほうが
良い検定であるといえるが、一般には単純な順序関係にはなっていない
ので、様々な作り方がありうる。

ただし、帰無仮説と対立仮説がともにひとつの確率分布のときには
最適な検定を定めることが可能である（次回）。

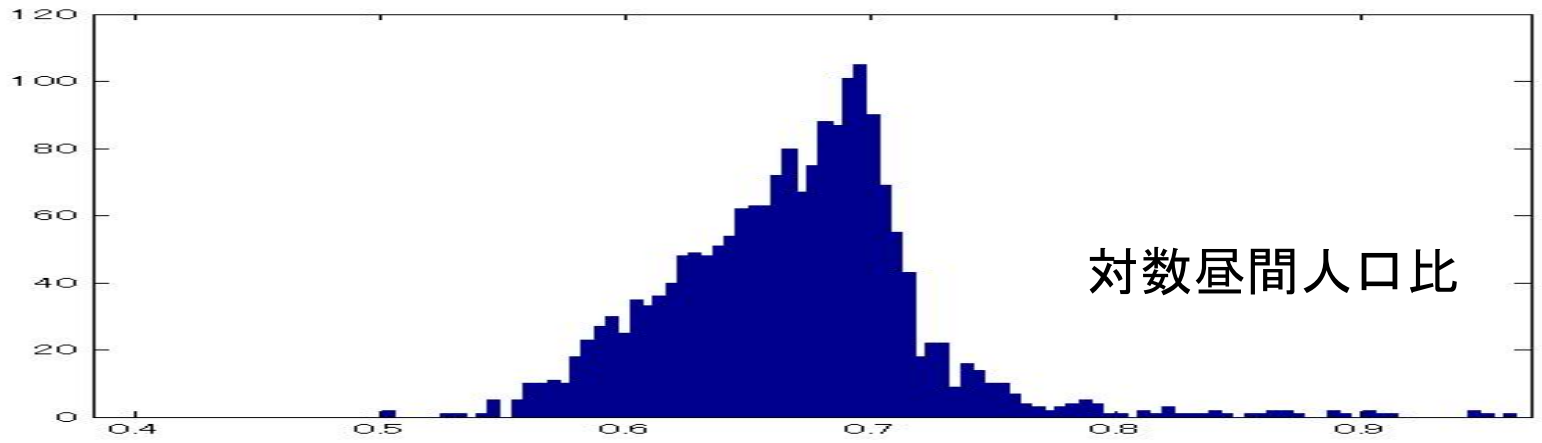
具体的な例

政府統計の総合窓口より http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

日本の市区町村 1901 個に対し対数昼間人口比 = log(1+昼間人口/人口)
と定義して、1901個の平均と標準偏差を求めたら 0.676, 0.097 だった。

都道府県47個について、第 k 番目の都道府県に含まれる市区町村を
n(k) とする。また 第 k番目の都道府県の第 m 番目の市区町村の
対数昼間人口比を X_{km} とする。

市区町村1901個の対数昼間人口比のヒストグラム
（正規分布とは言えませんが、今回は帰無仮説を正規分布としました。）



具体的な例

帰無仮説

「第 k 県の市区町村の対数昼間人口比は
平均 0.981, 標準偏差 0.443 の正規分布に従う」

対立仮説

「第 k 県の市区町村の対数昼間人口比は
平均 $\mu(\mu \neq 0.981)$ 標準偏差 σ の正規分布に従う」

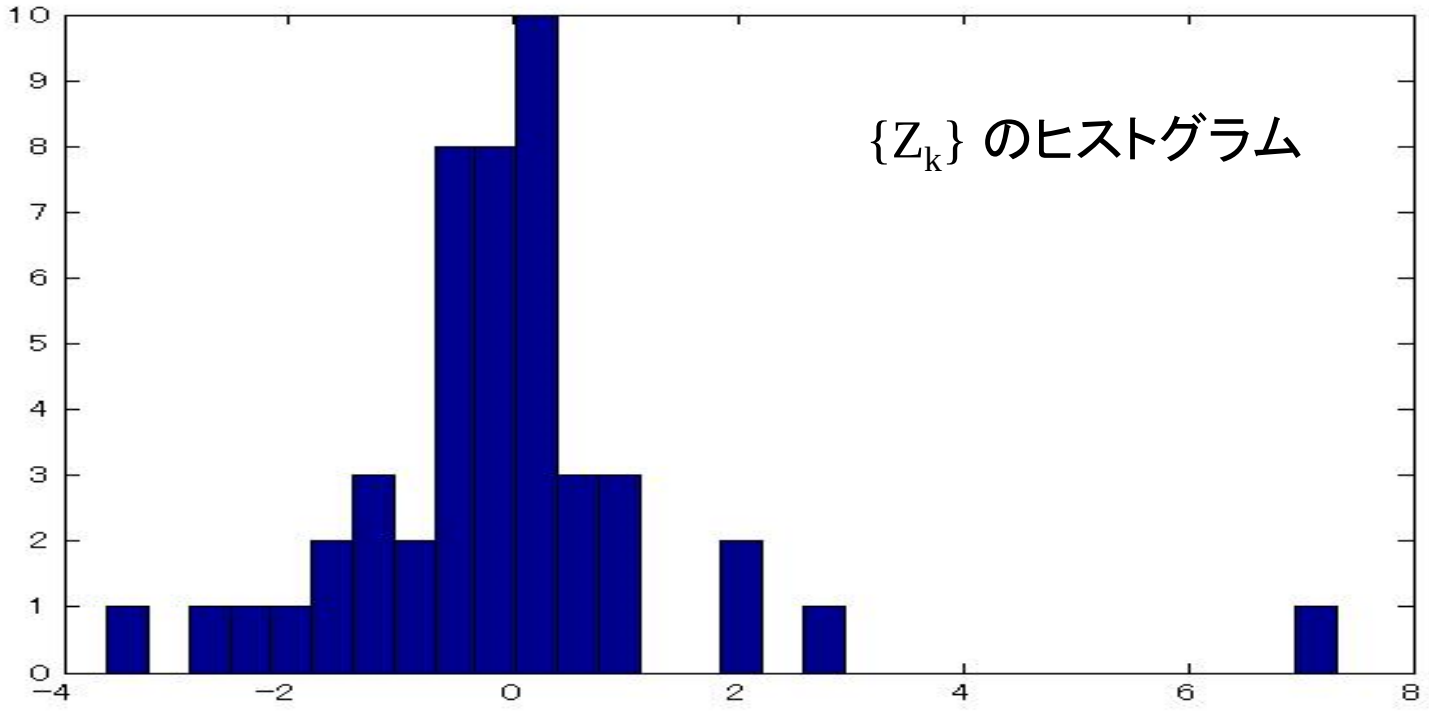
とする。帰無仮説が正しいとすると　第 k 県について

$$Z_k = \Sigma\,(X_{km} - 0.676) / (\,0.097\, n(k)^{1/2}\,)$$

は、平均0標準偏差1の正規分布に従う。$|Zk|>2.58$ のとき
帰無仮説は棄却される。

具体的な例

都道府県 47 について、Z_k の値をヒストグラムに書いた。
絶対値が2.58 を越えて、有意水準 0.01 で帰無仮説が棄却されたのは4都府県だった。東京都は明らかに他県とは異なることがわかる。



埼玉県 -3.61　千葉県 -2.65　　大阪府 2.64　　東京都 7.31