第一章　データの性質（補足　統計学の復習）

劉慶豊[1]

小樽商科大学

April 13, 2011

[1]E-mail:qliu@res.otaru-uc.ac.jp, URL:http://www.otaru-uc.ac.jp/~qliu/

確率変数

- 確率変数が取る値または取る値の範囲はおのおのに一定な確率が対応している。

離散確率変数

連続確率変数

実現値

ヒストグラムと確率密度関数（Probability Density Funtion PDF）

度数分布表

大学生男子50人の身長のデータ（DATA01）の度数分布

度数 各階級に入っているデータの数．相対度数：度数/全体のデータ数。

累積度数 下の階級からの度数の合計。相対累積度数：累積度数/全体のデータ数。

ヒストグラム

各階級の度数を棒グラフにしたものをヒストグラムという。

密度関数

正規分布を例

密度関数に関して

- 密度関数は面積で確率を表現する。
- 密度関数の曲線とX軸で囲まれた図形の全体の面積は1となる。
- 確率変数がある値以下になる確率はその値より左側の密度関数の曲線の下の面積に対応する。
- 確率変数が二つの値の間に入る確率はその二つの値の間の密度関数の曲線の下の面積に対応する。

正規分布の密度関数 期待値がμで分散がσ^2の正規分布の密度関数（通常小文字のxでXの特定の実現値を表す）

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

累積相対度数と累積分布関数

累積分布関数$F(x)$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx \tag{2}$$

密度関数と分布関数の関係

正規分布の累積分布関数

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

確率変数の標準化

期待値$E(X)=\mu$、分散$V(X)=\sigma^2$の確率変数Xがあるとする。Xから期待値を引いて、標準偏差で割って、できた新しい確率変数

$$Z=\frac{X-\mu}{\sigma}$$

は標準化された確率変数で、その期待値$E(Z)=0$、分散$V(Z)=1$となる。

分布表から確率を読み取る

- **期待値$\mu = 1$分散$\sigma^2 = 4$の正規分布確率変数Xに関して$X \leq 4.28$の確率を標準正規分布表（テキストp293、付表2）から読み取る。**

期待値の検定

以下の検定の話が成り立つための前提条件 標本数がかなり大きいまたは母集団が正規分布に従うとする。

仮説 帰無仮説H_0：棄却（否定）したい仮説。

対立仮説H_1：採択し（認め）たい仮説。

例 職業訓練の効果を調べたいとする。訓練を受けたと受けていない100人ずつの二つのグループの人の所得を調べる。

H_0：訓練を受けたグループの平均所得と受けていないグループの平均所得が同じ；
H_1：訓練を受けたグループの平均所得と受けていないグループの平均所得が異なる。

検定の根拠となる定理

標本数が少ない場合

Theorem

確率変数Xが正規分布に従うなら、その標本平均$\bar{X}$も正規分布に従う。

$$z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1).$$

ここでは「$\sim N(0,1)$」は「平均が0分散が1の正規分布に従う」の意味。
平均が0分散が1の正規分布は標準正規分布と呼ばれる。

検定の根拠となる定理（続き）

標本数が大きいとき

Theorem (中心極限定理)

独立同一な分布に従う確率変数の平均は，サンプル数が大きくなるに従いその期待値に近づく。すなわち、各X_iが平均（期待値）μと分散σ^2を持つ独立同一な分布に従うとき，nが大きくなるにつれ，$\sqrt{n}$倍した標準化した$\bar{X}$が標準正規分布に収束する（近づく）。

$$z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

σは普通未知であるため、σの代わりにσの推定量S（標準偏差、標本分散の平方根、以前お講義で勉強したS_x）を使う。

検定の根拠となる定理（続き）

Theorem

上の定理の条件の下で、σの代わりにσの推定量s（標準偏差、標本分散の平方根）を使った場合、t値

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \xrightarrow{d} t_{(n-1)}.$$

となる。ただし、$t_{(n-1)}$は自由度$n-1$のt分布を表す。

t分布

正規分布は平均と分散によって密度関数が決まるが、t分布は自由度というパラメーターにより密度関数が決まる。自由度が大きくになるにつれt分布が正規分布に近づき、密度関数のグラフが同じようになっていく。
下のグラフは標準正規分布の密度関数のグラフ（赤い点線）と自由度が3のt分布の密度関数のグラフ（ブルーの実線）

平均の片側検定

平均の検定には片側検定と両側検定がある、片側検定は平均がある値より大きいかどうか、または小さいかどうかに関して別々に検定する。両側検定の場合、両方を同時に検定する。

平均の片側検定の例題

国民の平均年収 が$20,000$ドルに達したかどうかが先進国であるかどうかを判断するための一つおおよそな指標となる。A国が先進国であるかどうかをこの指標で検討したいとする。A国の国民の個人年収を確率変数Xとする。Xの期待値（全国民の平均年収）に関して検定することを考える。10000人をくじ引きで選んで（無作為標本)、年収のデータを収集する、その平均を計算して$\bar{X} = 20,200$ドルになったとする。この10000人の年収の標本標準偏差を計算して$s = 7500$になったとする。期待値μが$20,000$ドルより大きいかどうかを検定する。

有意水準と臨界値（有意水準点）

- **有意水準を$a = 5\%$にした場合、影の部分の面積は$a = 5\%$を表している。その影の左端の座標は臨界値となる。今のグラフの例では臨界値は1.65となる。数式で表すと、$P\left(X > 1.65\right) = 5\%$。**

解答

- **有意水準を$\alpha = 5\%$とする。**
- **帰無仮説H_0：期待値$\mu \leq 20000$；対立仮説H_1：期待値$\mu > 20000$とする。**
- **検定統計量t値を計算する。**
  $t = \sqrt{n}\left(\bar{X} - \mu\right)/s = 100 \times (20200 - 20000)/7500 = 2.66$
- **t分布表より（自由度が大きい場合標準正規分布表を利用しても良い）、自由度（平均の検定のとき、自由度は$n-1$）が9999のt分布の5%の臨界値（有意水準点）が約1.65である。**
- **検定統計量の値$t = 2.66 > 1.65$であるため[2]帰無仮説H_0：期待値$\mu \leq 20000$が棄却される。国民の平均年収が$20,000$ドルを超えて先進国といえると判断する。**

[2] $\bar{X}$がその実現値20100を超える確率が5%よりも小さいと意味する

平均の両側検定

部品のサイズの平均の検定を例に説明する

工場から送ってきた大量な同じ種類の部品を検査することを考える。納品の中から100個を無作為に抽出して、直径を図り平均と標準偏差を計算し$\bar{X} = 3.2\,\text{cm}, s = 2$となった。良品の条件として直径の期待値が$\mu = 3$と決まっている。$\bar{X}$の値を利用して、$\mu = 3$であるかどうかの検定を考える。

方針 方法は片側検定とほぼ同じである。異なるところは

1. 帰無仮説は$=$を使う、対立仮説は$<, >$両方を使う。たとえば、H_0：$\mu = 3$；H_1：$\mu > 3$または$\mu < 3$。
2. 有意水準を決めてから、それを半分にして、有意水準点の値を調べる。たとえば、有意水準を5%にした場合、それの半分2.5%の有意水準点の値を調べる。
3. 検定統計量t値の計算などは片側のときと同じであるが、最後にtの絶対値と有意水準点の値（たとえば、2.5%の有意水準点の値）と比較する。$|t| >$有意水準点の値であれば、帰無仮説を棄却する、逆に$|t| \leq$有意水準点の値であれば、帰無仮説を採択する。

解答

- **両側検定を行う。有意水準を**5%**とする。**
- H_0**：**$\mu = 3$**；**H_1**：**$\mu > 3$**または**$\mu < 3$**。**
- $t = \sqrt{n}\left(\bar{X} - \mu\right)/s = 10 \times 0.2/2 = 1$**、自由度**$99$**の**$2.5\%$**有意水準点が**$1.96$**なので**$|t| < 1.96$**、帰無仮説は採択される。納品は良品であると判断する。**