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

October 13, 2009

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

確率変数

- 確率変数が取る値はおのおのに一定な確率が対応している。
- コイン投げの実験では表を1とし裏を0としたら、コイン投げの結果を確率変数Xのとる値とそれに対応している確率は
  $X = 0$の確率は0.5, $X = 1$の確率も0.5である。
- サイコロの場合は$X = 1$の確率は$1/6$, $X = 2$の確率も$1/6$, ...。

離散確率変数

- 上述した二つの確率変数の例では確率変数Xは0と1または1, 2, 3, 4, 5, 6と飛び飛びとした値しかとらない、たとえば0と1の間の値を取ることがない、この場合は離散確率変数と呼ぶ。
- 確率変数が取る値とそれに対応する確率との関係は以下の表で表せる。

Xが取る値	x_1	x_2	$\cdots$	x_n
対応する確率$P(x)$	$P(x_1)$	$P(x_2)$	$\cdots$	$P(x_n)$

サイコロの例

Xが取る値	$x_1 = 1$	$x_2 = 2$	$\cdots$	$x_6 = 6$
対応する確率$P(x)$	$P(x_1) = 1/6$	$P(x_2) = 1/6$	$\cdots$	$P(x_6) = 1/6$

度数関数、分布関数

確率度数関数 ここでは$P(x)$は確率度数関数と呼ばれる。
$P(x_1), P(x_2) \ldots$はxの具体的な値に対応する確率を表す。

分布関数 分布関数の定義は$F_X(c) = P(\{X \leq c\})$となる。表しているのはXがcより小さくなる確率である。離散確率変数の場合は

$$F_X(c) = P(\{X \leq c\}) = \sum_{x_i < c} P(x_i)$$

サイコロの例

$$\begin{aligned} F_X(3) &= P(\{X \leq 3\}) = \sum_{x_i < c} P(x_i) \\ &= P(x_1) + P(x_2) + P(x_3) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \end{aligned}$$

確率変数の期待値

確率変数Xの期待値を$E(X)$で表す。期待値は直感的に確率変数の理論上の平均で理解できる。離散確率変数の場合

$$E(X) = \sum_{i=1}^{n} x_i P(x_i).$$

サイコロの例

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} x_i P(x_i) \\ &= P(x_1) \times x_1 + P(x_2) \times x_2 + \cdots + P(x_6) \times x_6 \\ &= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} \cdots + 6 \times \frac{1}{6} = \frac{7}{2} \end{aligned}$$

確率変数の分散

確率変数Xの分散を$V(X)$で表して、離散確率変数の場合

$$V(X) = E\left[(X - E(X))^2\right] = \sum_{i=1}^{n}\left[(x_i - E(X))^2 P(x_i)\right].$$

サイコロの例

$$\begin{aligned} V(X) &= \sum_{i=1}^{n}\left[(x_i - E(X))^2 P(x_i)\right] \\ &= (x_1 - E(X))^2 \times P(x_1) + \cdots + (x_6 - E(X))^2 \times P(x_6) \\ &= \left(1 - \frac{7}{2}\right)^2 \times \frac{1}{6} + \cdots + \left(6 - \frac{7}{2}\right) \times \frac{1}{6} \end{aligned}$$

確率分布

Definition (確率分布)

ある確率変数Xの取る値に対応する確率がある関数$P(x)$で表せるとき、この確率変数Xの確率分布は$P(x)$に従うという。

ベルヌーイ分布の例

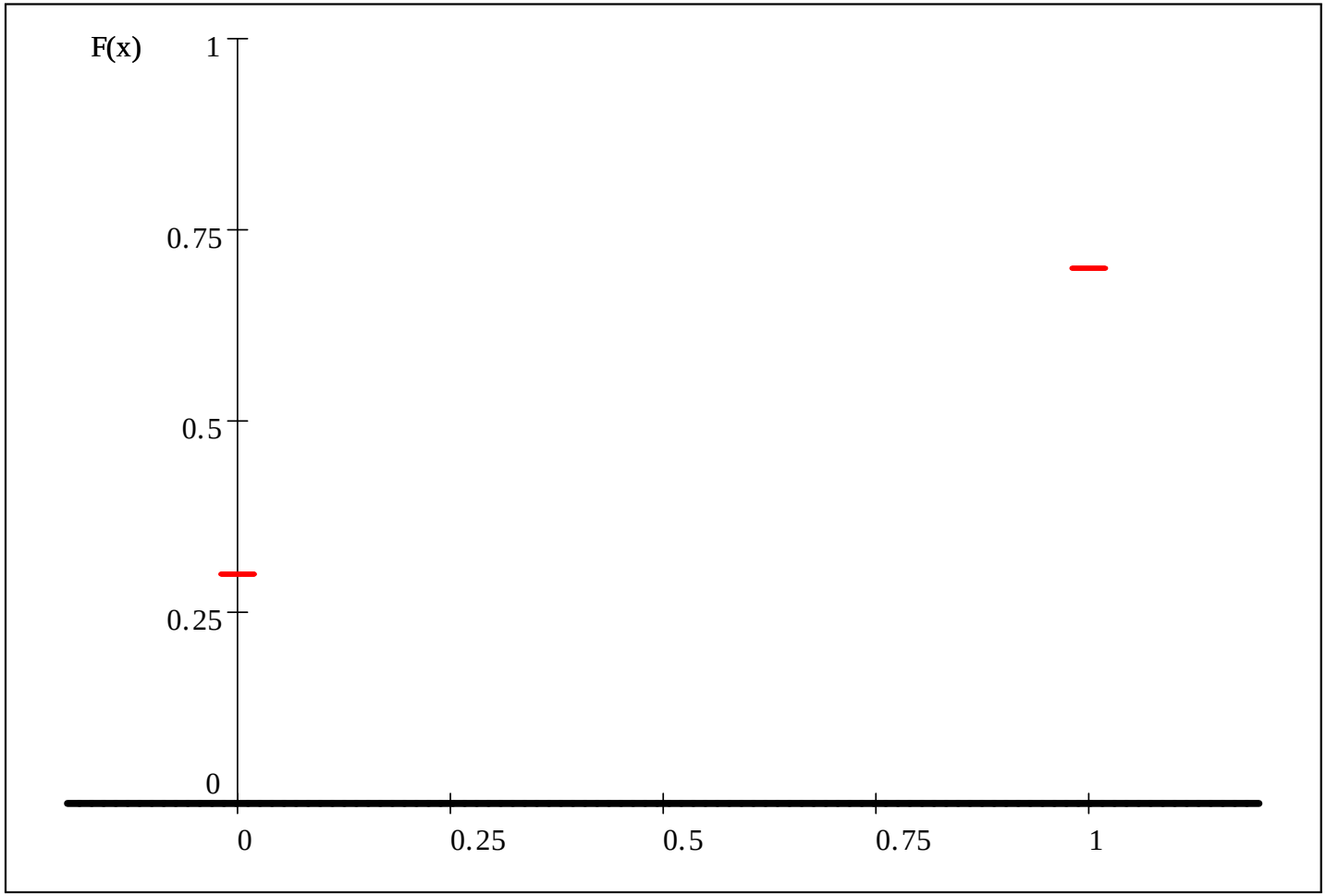
- **ベルヌーイ分布の確率度数関数**

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

  $p = 0.5$のときは１回コイン投げた時の結果の分布となる。
- **期待値**$E(X) = 1 \times p + 0 \times (1 - p) = p$
- **分散は**

$$\begin{aligned} V(X) &= (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p \\ &= p^2 - p^3 + p - 2p^2 + p^3 = p - p^2 \end{aligned}$$

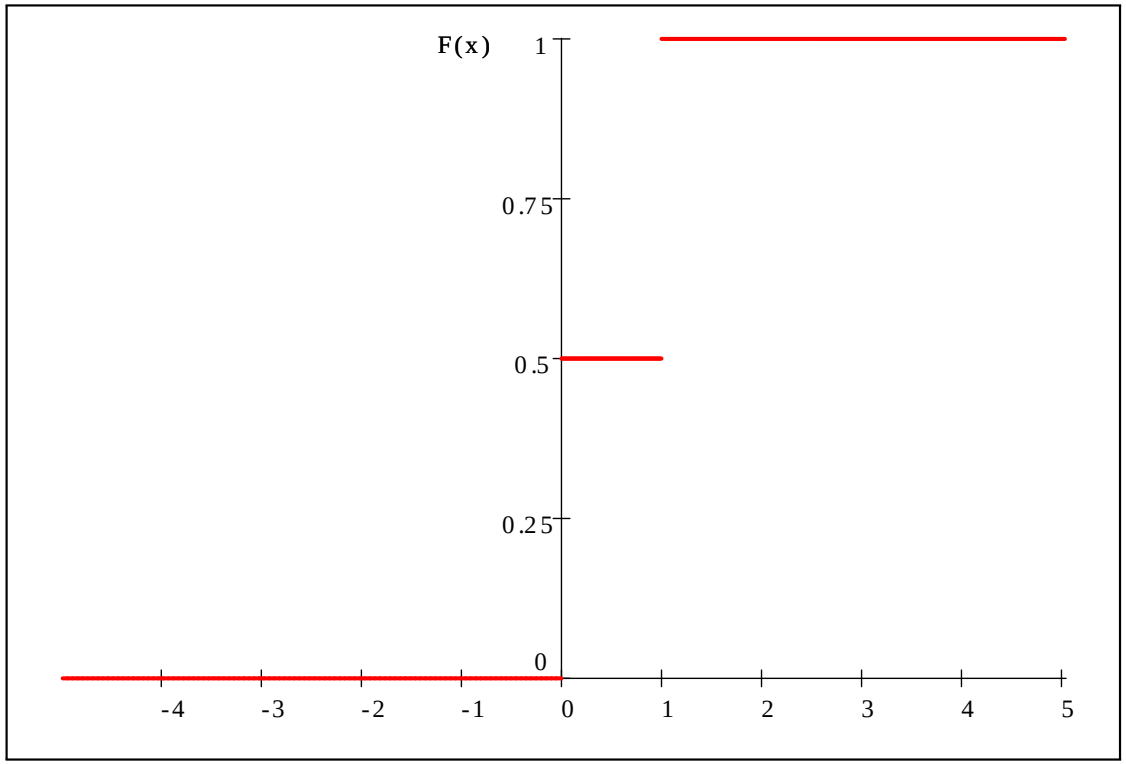
- **確率度数関数のグラフ**



ベルヌーイの確率度数関数　$p = 0.7$

- **確率分布関数**

$$F(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$



ベルヌーイ分布の分布関数 $p = 0.5$

連続確率変数

連続確率変数 連続な値をとる確率変数。たとえば、部品の誤差、新生児の身長。

密度関数 連続確率変数の度数関数は密度関数と呼ぶ。

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

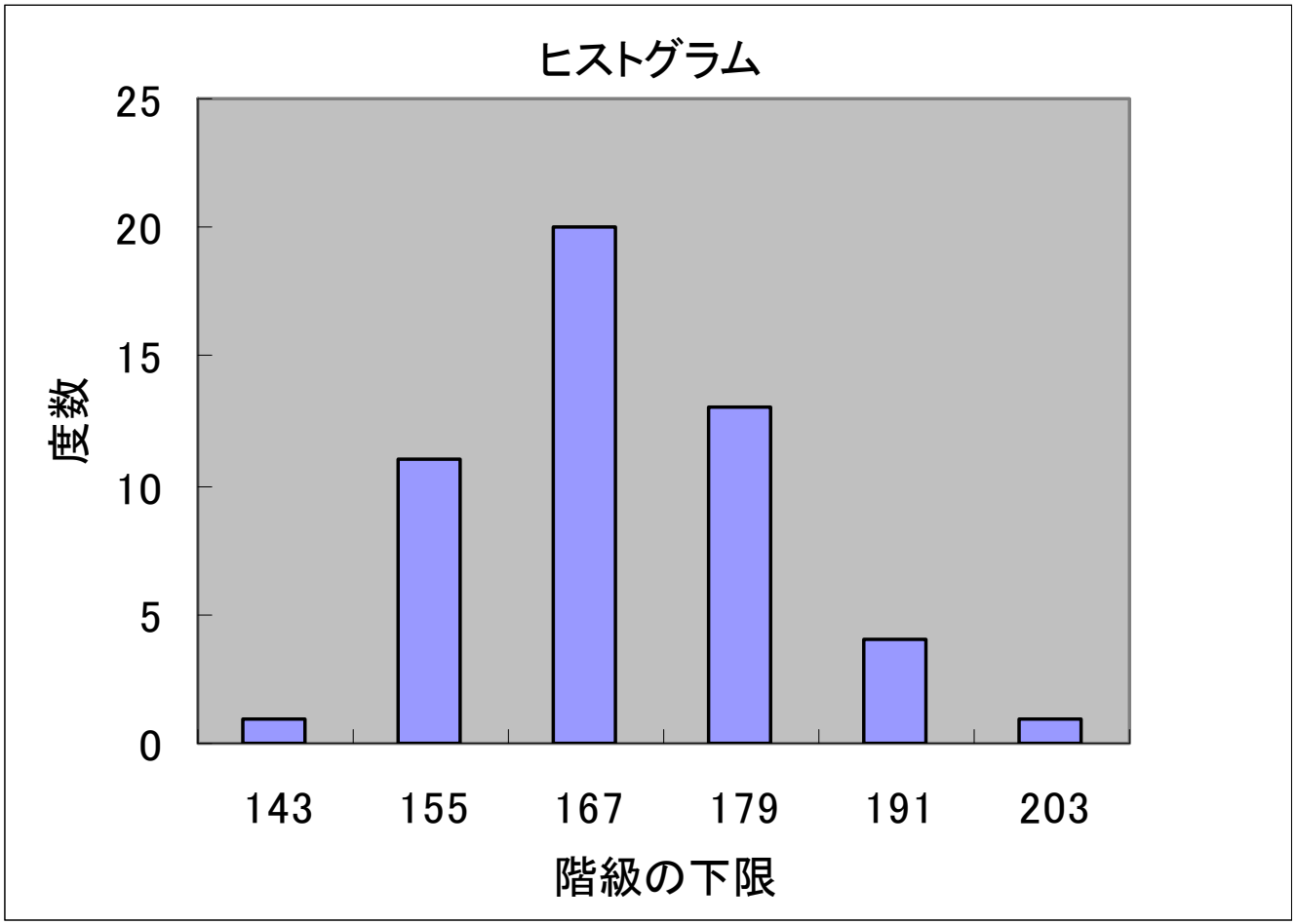
階級	度数	累積度数	相対度数	累積相対度数
143–152	9	9	18%	18%
152–161	10	19	20%	38%
161–170	14	33	28%	66%
170–179	12	45	24%	90%
179–188	2	47	4%	94%
188–198	3	50	6%	100%

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。

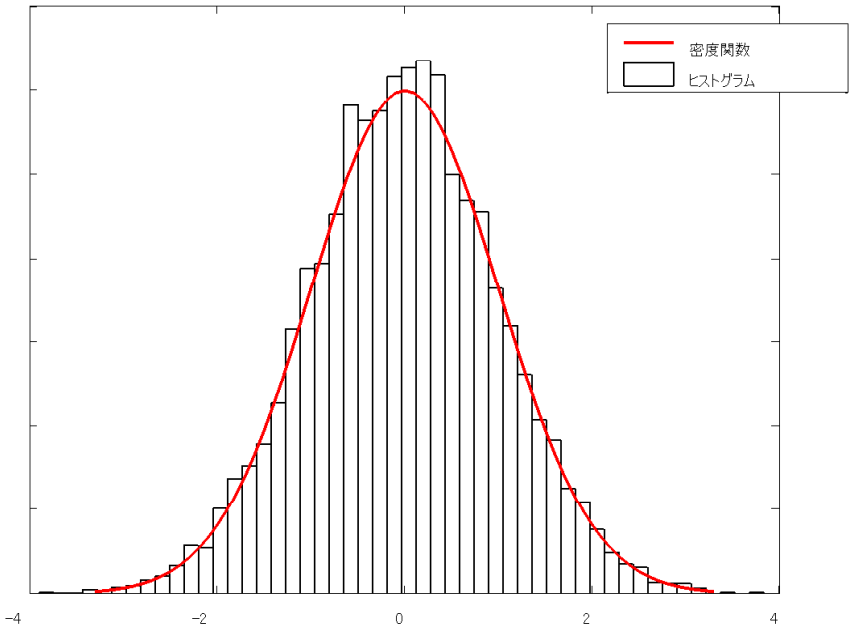
ヒストグラム

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

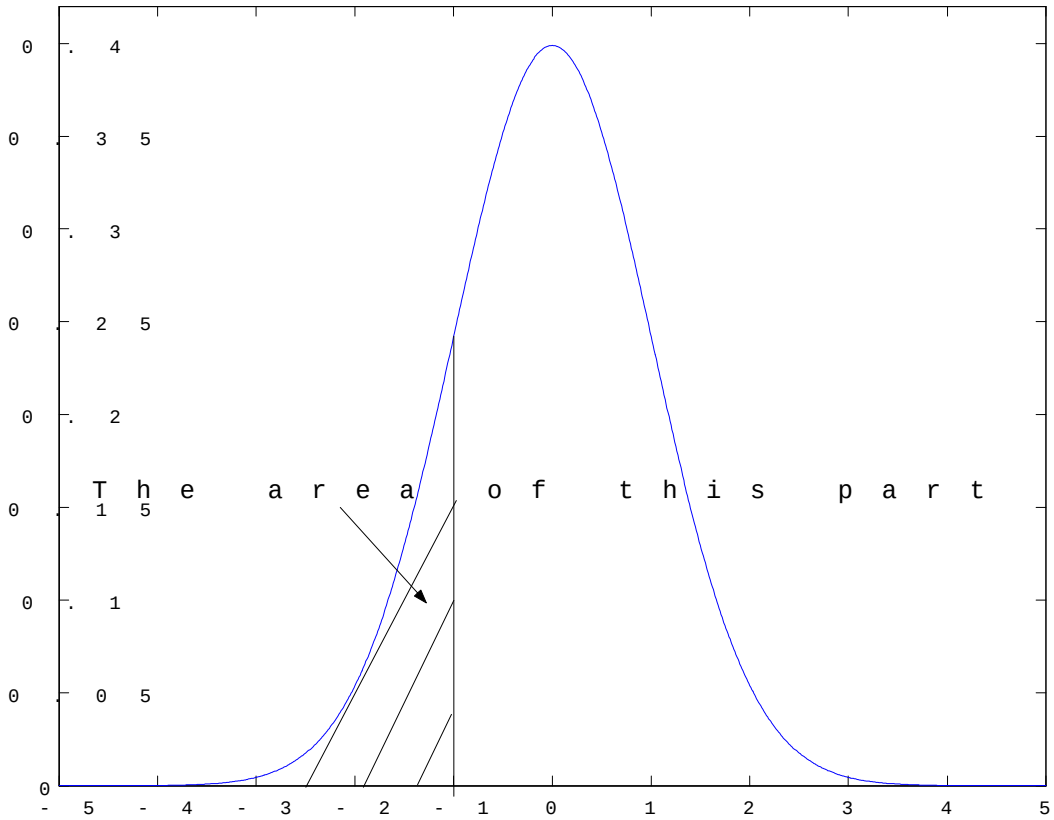
正規分布を例 標準化した身長のデータのヒストグラム（相対度数で描いたもの）、その上に標準正規分布の密度関数を重ね合わせた。正規分布の密度関数のグラフは鐘または富士山の形をしている。



正規分布

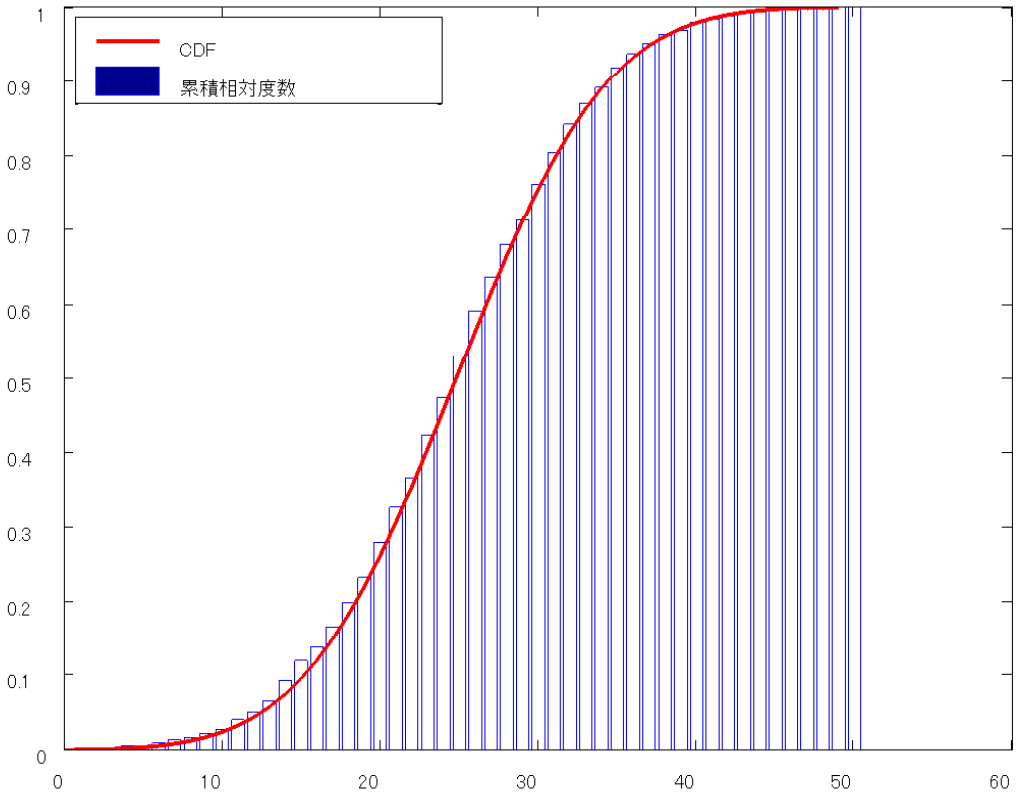
正規分布の確率密度関数

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



-1以下の値になる確率

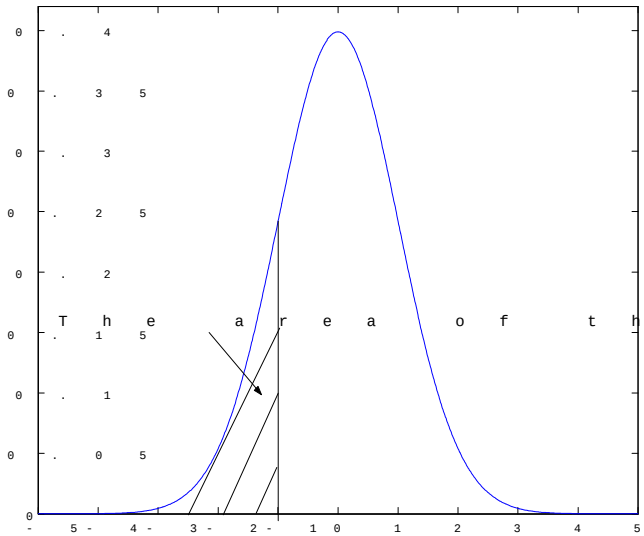
累積相対度数と累積分布関数



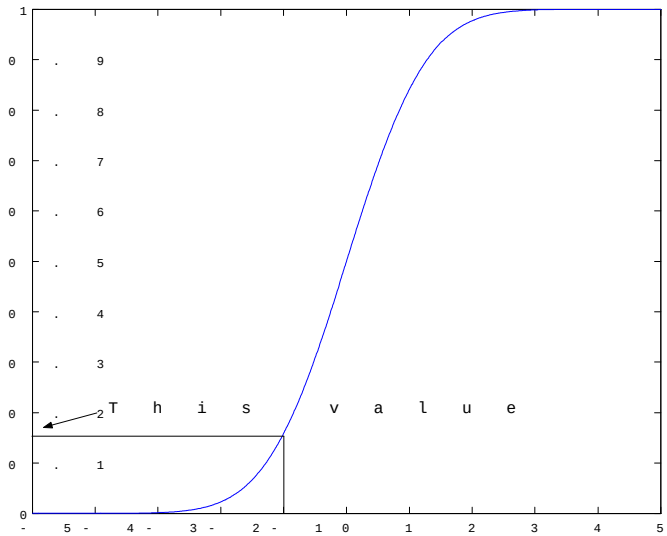
累積分布関数$F(x)$ 略して分布関数、ある値cに対応する累積分布関数の値はcより小さい値（cを含む）に対応する確率の総和である。

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

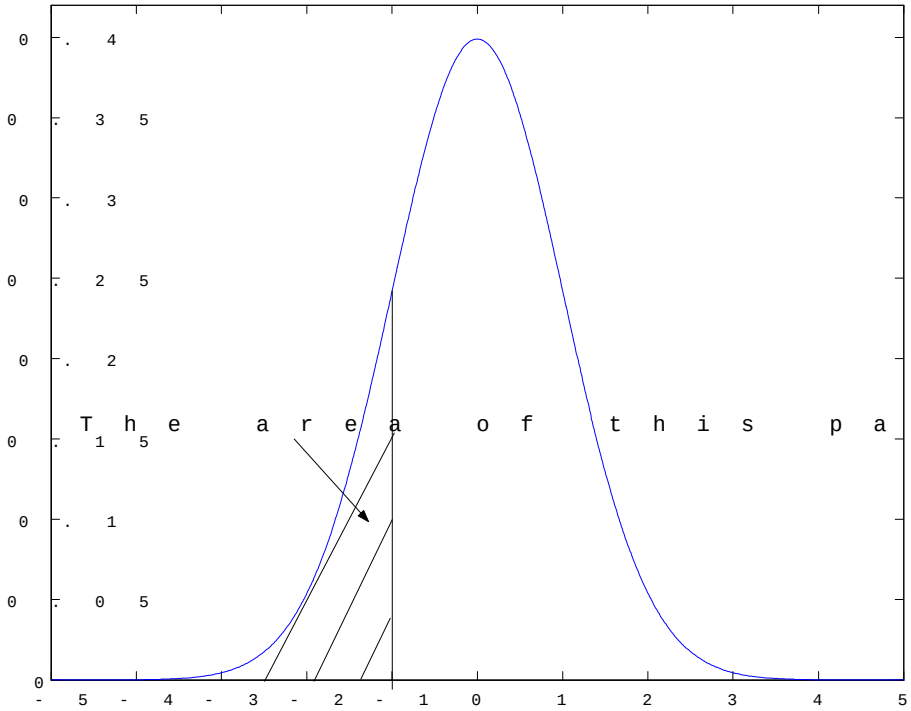
密度関数と分布関数の関係 分布関数は確率密度関数の積分の形になっている。積分が図形の面積の計算に対応していることから、連続確率変数の場合に関して、標準正規分布を例にグラフに描けば以下のようになる：



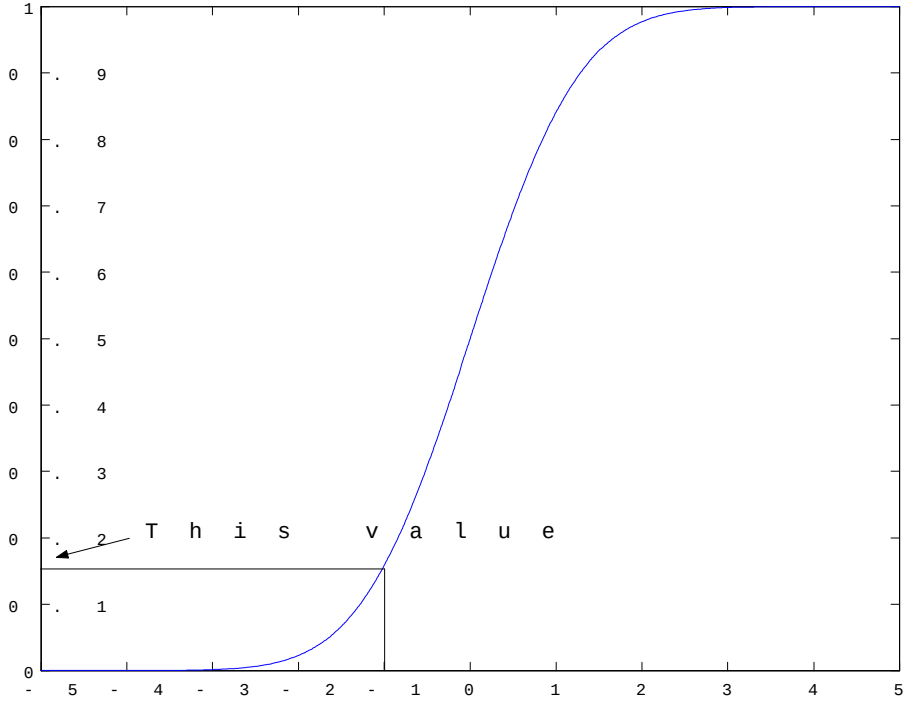
$F(-1)$の値（密度関数で表す）



$F(-1)$の値（分布関数で表す）



$F(-1)$の値（密度関数で表す）



$F(-1)$の値（分布関数で表す）

正規分布の累積分布関数

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

期待値と分散

前回は既に離散確率変数の期待値と分散に関して説明した。離散確率変数の場合

$$E(X) = \sum_{i=1}^{n} x_i P(x_i).$$

確率変数の分散：確率変数Xの分散を$V(X)$で表して、離散確率変数の場合

$$V(X) = E\left[(X - E(X))^2\right] = \sum_{i=1}^{n} \left[(x_i - E(X))^2 P(x_i)\right].$$

連続関数の場合は

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx.$$

$$V(X) = E\left[(X - E(X))^2\right] = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx.$$

正規分布のまとめ

- **正規分布の確率密度関数**

$$f(c) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(c-\mu)^2}{2\sigma^2}} \tag{3}$$

- **正規分布の累積分布関数**

$$F(c) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{4}$$

- **正規分布の期待値はμ分散がσ^2：**

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \mu \\ V(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \sigma^2 \end{aligned}$$

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。**
- **まずXを標準化してZとする。Zが標準正規分布（期待値$\mu = 0$分散$\sigma^2 = 1$）に従う確率変数となる。**

$$Z = \frac{X - \mu}{\sqrt{\sigma^2}} = \frac{X - 1}{2}$$

- **Zに関して標準正規分布表を適用する。$x = 4.38$のとき（小文字のxでXの実現値を表す）**

$$z = \frac{x - 1}{2} = \frac{4.28 - 1}{2} = 1.64$$

ゆえに、 $P(X \leq 4.28) = P(Z \leq 1.64) \approx 0.95$**。**

期待値と分散の推定

推定には点推定と区間推定があるが、本講義では点推定だけ説明する。

母集団 分析対象の全体である。標本を母集団の中から抽出される。たとえば、日本の国民の所得を分析したい場合、全部の国民の所得が母集団を構成する。世界全人口の所得を対象に分析したい場合、地球上にいるすべての人の所得が母集団を構成する。

無作為標本 母集団のすべての個体が均等な機会でかつ互いに無関係（独立）に抽出されるように抽出された標本（公平なくじ引きで考えれば分かりやすい）。

期待値と分散の推定（続き）

- 確率変数Xの実現値$\{X_1, X_2, X_3, \ldots, X_n\}$を$X$の分布に従う母集団からの無作為標本とする。

母集団の期待値$\mu = E(X)$の推定方法

無作為標本の算術平均（標本平均）$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$を用いて推定する。

母集団の分散$\sigma^2 = E(X-\mu)^2$の推定方法

一つの推定量として標本分散$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$用いる。

期待値（時には平均値とも呼ばれる）と分散の推定例

確率変数X 商大生の身長を確率変数と見なす。

母集団 商大の学生全員の身長。

無作為標本 商大の学生全員のくじを作り、くじ引きで100人を選び、身長を測って無作為標本$\{X_1, X_2, X_3, \ldots, X_{100}\}$とする。

$E(X)$またはμ 商大生全員の平均身長、母集団の期待値（平均）。

μの推定量

$$\bar{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$$

σ^2の推定量

$$S^2 = \frac{1}{100-1}\sum_{i=1}^{100}(X_i - \bar{X})^2$$

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。

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

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）。

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

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）を使う。

検定の根拠となる定理（続き）

Theorem

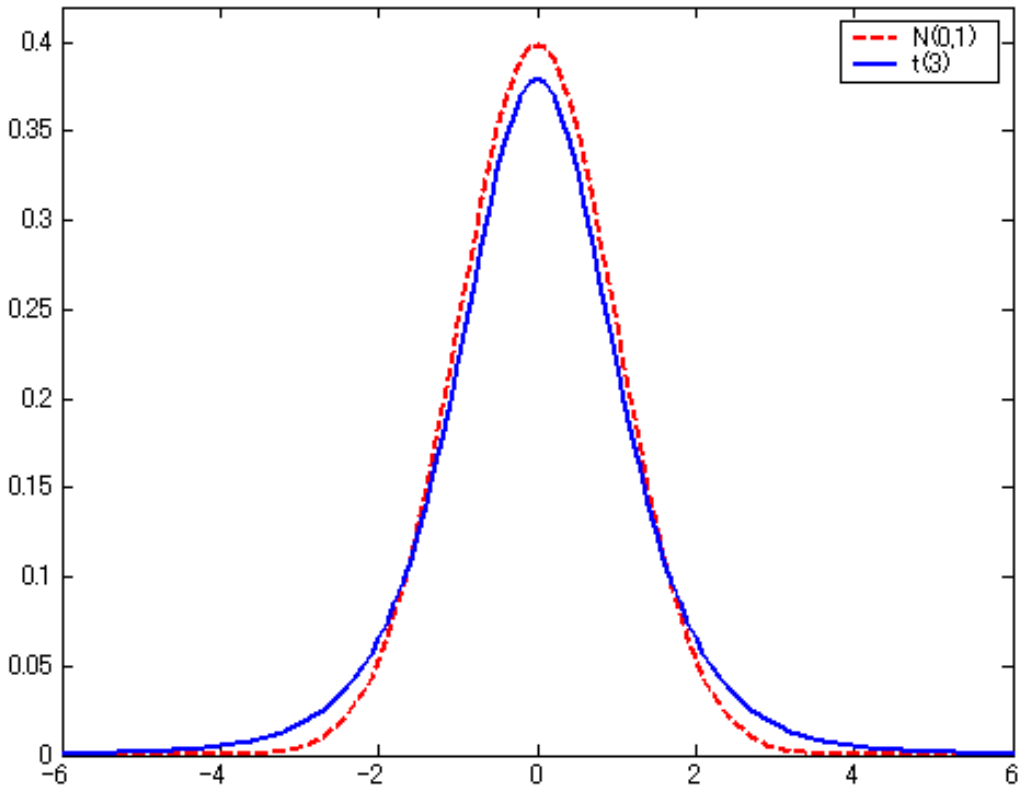
上の定理の条件の下で、σの代わりにσの推定量s（標準偏差、標本分散の平方根）を使った場合

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

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$\bar{X} = 20,100$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。

基本的な考え方

- まずXは平均（期待値）が$\mu = 20000$、標準偏差が$\sigma = s = 7500$の正規分布$N(20000, 7500)$に従うと仮定する。
- 上述の仮定の下で、$\bar{X}$が実現値20100を超える確率は極めて小さいなら、仮定が$\bar{X}$の実現値20100、すなわちデータと矛盾することを意味する。
- そのため、もし、この確率がとても小さく、事前に決めた許容値α（有意水準と呼ばれる、通常1%か5%とする）よりも小さいなら、$\mu \leq 20000$の仮説が偽であると判断する（棄却する）。
- 逆の場合、$\mu \leq 20000$の仮説が真であると判断する（採択する）。

解答

- **有意水準を**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**$t = \sqrt{n}\left(\bar{X} - \mu\right)/s = 100 \times (20200 - 20000)/7500 = 2.66$
- t**分布表より、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。**

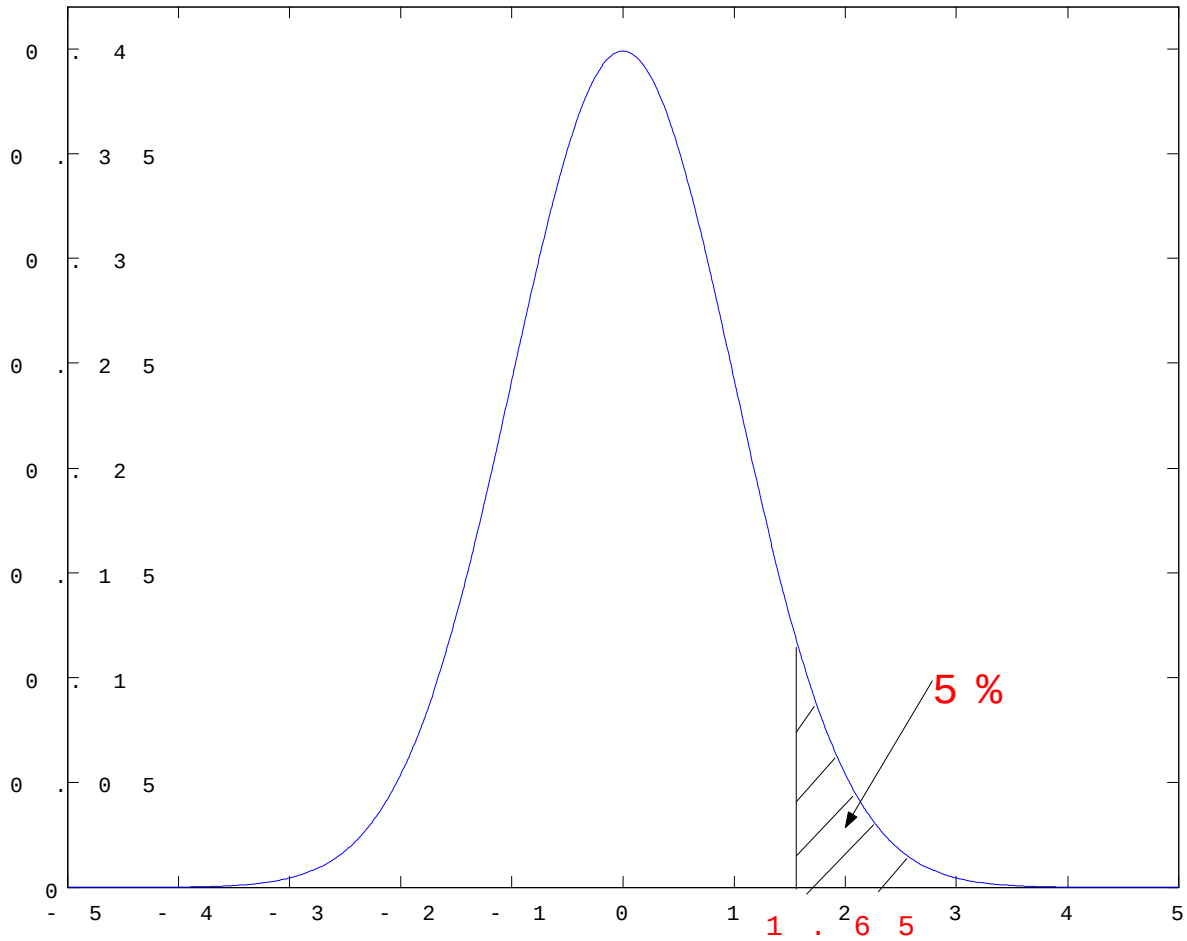
[2] $\bar{X}$**がその実現値**20100**を超える確率が**5%**よりも小さいと意味する**

注

1. 今の例題の対立仮説は$\mu > 20000$、より大きい$(>)$となっている。より小さい（$<$）たとえば対立仮説が$\mu < 20000$となる場合、t値を計算して左側の有意水準点と比較する。t値が左側の有意水準点より小さければ帰無仮説を棄却する、逆の場合だったら帰無仮説を採択する。
2. 左側の有意水準点は分布表から見つかった右側の有意水準点の値$\times(-1)$)、たとえば、右側の5%有意水準点が1.65だったら、左側は-1.65となる。

密度関数のグラフで見る

グラフで示すなら、検定統計量 t が t 分布の裾の端に入って有意水準点よりも右にあって、$\mu \leq 20000$ の仮説が偽であると判断し棄却する。



解答に至るまでのステップ

1. 有意水準を選ぶ。たとえば、有意水準を $\alpha = 5\%$ とする。（普通は 1% か 5% にする）
2. 仮説を立てる。帰無仮説 H_0：期待値 $\mu \leq 20000$；対立仮説 H_1：期待値 $\mu > 20000$。
3. t 分布の 5% 有意水準点の値を読み取る。
   例の場合約 1.65。
4. $\bar{X}$、s を計算して統計量 $t = \sqrt{n}\left(\bar{X} - \mu\right)/s$ を計算する。
   $t = 100 \times (20200 - 20000)/7500 = 2.66$。
5. ステップ３で求めた t とステップ２で求めた有意水準点の値 1.65 と比較する、$t > 1.65$ であれば、帰無仮説を棄却する。$t \leq 1.65$ であれば帰無仮説が採択される。例では $t > 1.65$ なので、帰無仮説を棄却する。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。

演習問題

ある美容室が割引サービスを行った、この割引サービスによって、一日の平均来客数が増えたかどうかを調べたい。この美容室の普段の平均来客数が10人。割引サービスを実施後、25日間来客数を集計して平均と標準偏差を計算して$\bar{X} = 12$、$s = 3$だとする。検定を行って一日平均の来客数が増えたかどうかを判断してください。ヒント：H_0：$\mu \leq 10$；H_1：$\mu > 10$。

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。**