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

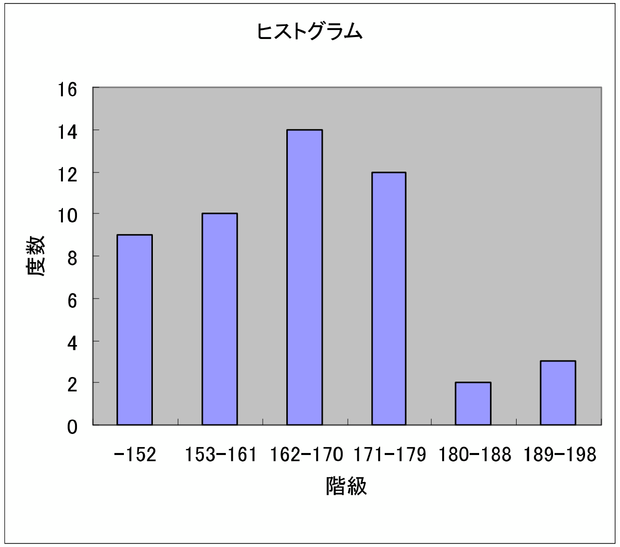
階級	度数	累積度数	相対度数	累積相対度数
143–152	9	9	18%	18%
152–161	10	19	20%	38%
161–170	14	33	28%	66%
170–179	12	45	24%	90%
179–188	2	47	4%	94%
188–198	3	50	6%	100%

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。

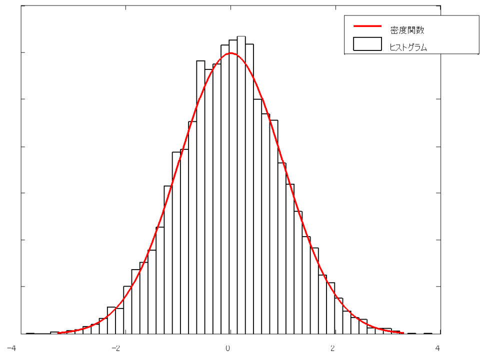
ヒストグラム

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

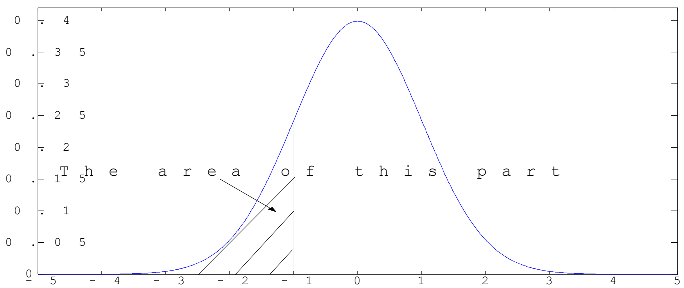
正規分布を例に 標準化した身長のデータのヒストグラム（相対度数で描いたもの)、その上に標準正規分布の密度関数を重ね合わせた。正規分布の密度関数のグラフは鐘または富士山の形をしている。



正規分布

密度関数に関して

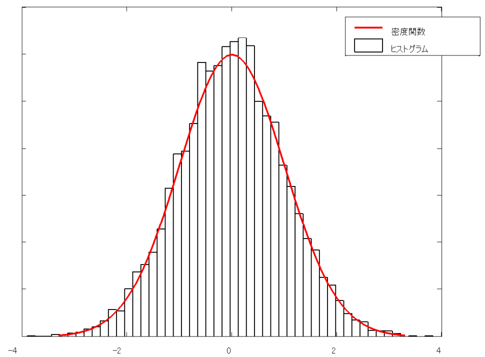
- 密度関数は面積で確率を表現する。
- 密度関数の曲線と X 軸で囲まれた図形の全体の面積は1となる。
- 確率変数がある値以下になる確率はその値より左側の密度関数の曲線の下の面積に対応する。
- 確率変数が二つの値の間に入る確率はその二つの値の間の密度関数の曲線の下の面積に対応する。



-1以下の値になる確率

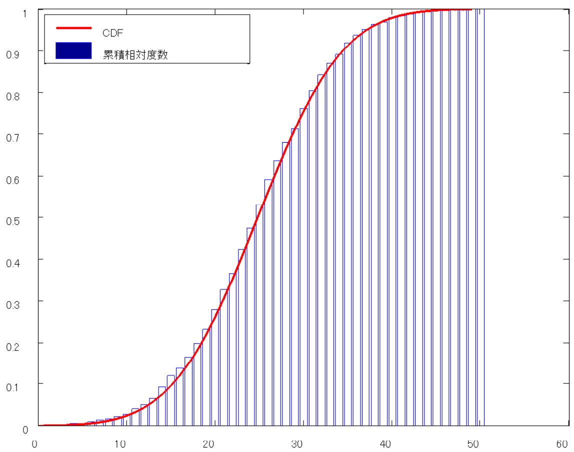
正規分布の密度関数 期待値が μ で分散が σ^2 の正規分布の密度関数（通常小文字の x で X の特定の実現値を表す）

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



正規分布

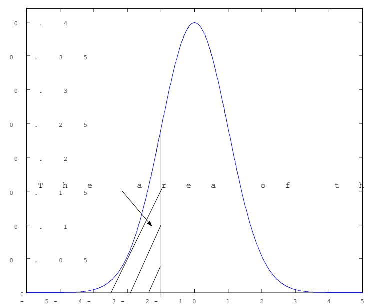
累積相対度数と 累積分布関数



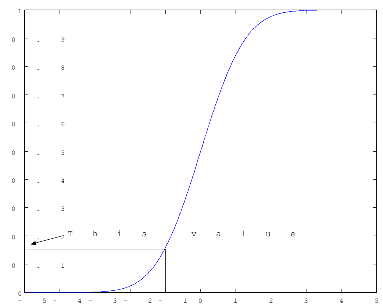
累積分布関数$F(x)$ 略して分布関数、確率変数Xがある特定の値xに対応する累積分布関数の値はxより小さい値（xを含む）に対応する確率の総和である。

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

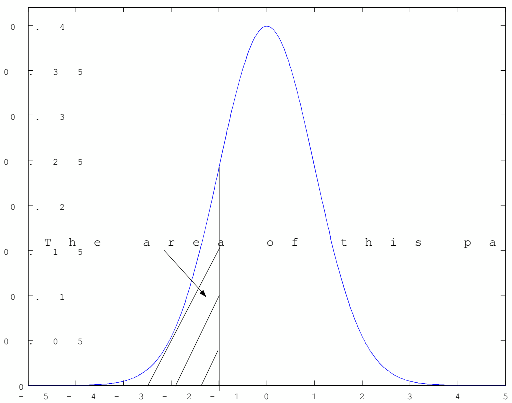
密度関数と 分布関数の 関係 分布関数は確率密度関数の積分の形になっている。積分が図形の面積の計算に対応していることから、連続確率変数の場合に関して、標準正規分布を例にグラフに描けば以下のようになる：



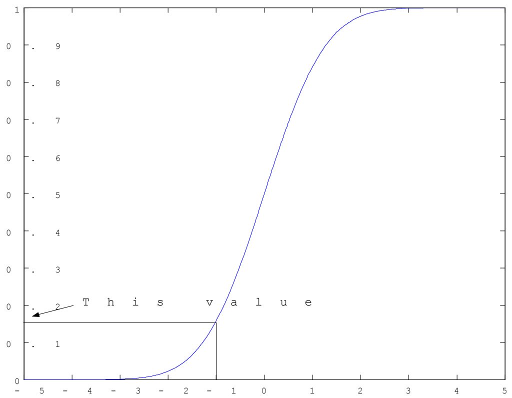
$F(-1)$の値（密度関数



$F(-1)$の値（分布関数で



$F(-1)$ 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

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

期待値と 分散

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

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

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

性質 $E(cX + b) = cE(X) + b$、$V(cX + b) = c^2 V(X)$、ただし cと bは定数である。

連続確率変数の 共分散と 相関係数

共分散の 定義

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))].$$

独立な 確率変数の 共分散 確率変数Xと Yが 独立であると き 共分散 $Cov(X, Y) = 0$。

確率変数の 相関係数

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)\,V(Y)}}$$

$$-1 \leq \rho(X, Y) \leq 1$$

正規分布のまとめ

- 正規分布は統計学の中で最もよく利用される分布の一つである。測量データ、誤差、平均値の分布、$\cdots$。
- 正規分布の確率密度関数

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

- 正規分布の累積分布関数

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

- 正規分布の期待値はμ分散がσ^2：

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

正規分布の再生性

正規分布の再生性

1. 確率変数X_i、 $i = 1, 2, \cdots, m$,が $N\left(\mu_i, \sigma_i^2\right)$に従って独立に分布しているとき、$\sum_{i=1}^{m} c_i X_i$は $N\left(\sum_{i=1}^{m} \mu_i, \sum_{i=1}^{m} c_i^2\sigma_i^2\right)$に従って分布する。ただし、$c_i$、$i = 1, 2, \cdots, m$,が定数の列である。
2. $i = 1$の場合、すなわちXが $N\left(\mu, \sigma^2\right)$のとき、$cX$の分布は $N\left(c\mu, c^2\sigma^2\right)$になる。ここでは $N\left(\mu, \sigma^2\right)$は「平均が$\mu$分散が$\sigma^2$の正規分布」を表す。

例　$X \sim N(1, 9)$、$Y \sim N(2, 1)$のとき、$2X \sim N(2, 36)$となり、$(2X + 3Y) \sim N(8, 45)$となる。ここでは「$\sim N\left(\mu, \sigma^2\right)$」は「平均が$\mu$分散が$\sigma^2$の正規分布に従う」の意味。

その他の連続分布（参考）

密度関数は区間 $[a, b]$ の中で一定で、それ以外はゼロである。密度関数の面積が 1 になるため、$a \leq x \leq b$ に関しては $f(x) = \frac{1}{b-a}$ となる。

性質 $E(X) = \frac{a+b}{2}$、$V(X) = \frac{(b-a)^2}{12}$

例 丸い鉛筆の横断面の長さが c とする。その横断面の縁の一点にマックを付けてを転がす。接地点とマックの間の弦の長さの分布が $[0, c]$ における一様分布である。

その他の連続分布（参考）

指数分布 密度関数は $x > 0$ に関して $f(x) = \lambda \exp(-\lambda x)$、$x \leq 0$ に関して $f(x) = 0$。布関数は $x > 0$ に関して $F(x) = 1 - \exp(-\lambda x)$、$x \leq 0$ に関して $F(x) = 0$。ただし、$\lambda > 0$ である。

例 失業者が再就職までの時間の分布の近似としてよく利用される。次の事故が起きるまでの時間の分布など。

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$X\le 4.28$の確率を標準正規分布表（テキストp279、付表4）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。$x=4.28$のとき

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

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

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$\{X_1, X_2, X_3, ..., X_n\}$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 $\{X_1, X_2, X_3, ..., X_{100}\}$ 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

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

推定量の性質

母集団のパラメーター（母数）を θ とし、推定量を $\hat{\theta}$ と表記する。

不偏性 推定量の期待値が推定したい母数と等しいとき、すなわち $E(\hat{\theta}) = \theta$、推定量が不偏であるという。

標本平均の不偏性 標本は無作為標本とする、すなわち、各 X_i の間は独立である。明らかに $E(\bar{X}) = \mu$、なので $\bar{X}$ が μ の不偏推定量である。

効率性（参考） 同じ母数の二つの推定量 θ_1 と θ_2 に関して、θ_1 の漸近分散が θ_2 のより小さい場合、θ_1 は θ_2 より効率的という。

標本平均の不偏性一致性

確率収束 確率変数Xと実数cは任意の正の実数δに関して

$$P(|X-c| \geq \delta) \to 0$$

が満たされるとき、確率変数Xが実数cに確率収束するという。$X \xrightarrow{p} c$と表記する。

一致性 サンプルサイズnが無限大に近づくにつれ、推定量が母数に確率収束する、すなわち$\hat{\theta} \xrightarrow{p} \theta$とき、推定量は一致性を持つという。

チェビシェフの不等式 任意の正の実数δに関して

$$P(|X-\mu| \geq \delta) \leq \frac{Var(X)}{\delta^2}$$

$\bar{X}$の分散 $Var(\bar{X}) = E\left[(\bar{X}-\mu)^2\right] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \sigma^2/n$、(独立の確率変数間の共分散が0であることを利用すれば証明できる。)

標本平均の一致性

標本平均の一致性 チェビシェフの不等式より任意の正の実数δに関して

$$P(|\bar{X}-\mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$

nが無限大に近づくにつれ、右辺は0に収束する。従って、$\bar{X} \xrightarrow{p} \mu$、一致性を持つ。

大数の法則 標本サイズnが無限になるにつれ無作為標本の標本平均$\bar{X}$は母集団平均μに確率収束する。すなわち$\bar{X} \xrightarrow{p} \mu$を満たす。

標本分散の分布

カイ２乗分布 Z_i、$i = 1, 2, \cdots, k$、が標準正規分布に従うとき、$W = Z_1^2 + Z_2^2 + \cdots + Z_k^2$が自由度$k$の$\chi^2$分布に従う。テキスト図4.5。

標本分散の分布に関する定理 正規分布$N(\mu, \sigma^2)$からの無作為標本の標本分散をS^2とする。

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

が自由度$(n-1)$のχ^2分布に従う。証明略。

自由度に関して（参考） $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$が必ず満たされるという制約がかかるため、自由度が1が減り、$(n-1)$となる。

期待値（母平均）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$\mathbf{H}_0$：棄却（否定）したい仮説。

対立仮説$\mathbf{H}_1$：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。

標本平均の分布１.

標本数が少ない場合

Theorem (検定の根拠となる)

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。

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

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。

標本平均の分布２.

標本数が大きいとき

Theorem (中心極限定理、検定の根拠となる)

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）。

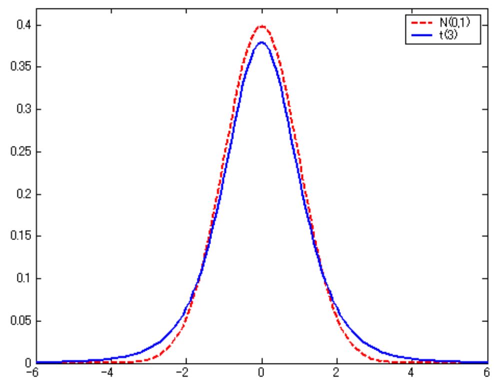
$$z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

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

t分布

t 分布 Zが標準正規分布、Wが自由度kのχ^2分布に従うとする。$t = \frac{Z}{\sqrt{W/k}}$がt分布に従う。

性質 正規分布は平均と分散によって密度関数が決まるが、t分布は自由度によって決まる。自由度が大きくになるにつれ正規分布に近づき、密度関数のグラフが同じようになっていく。下記グラフは標準正規分布と自由度が3のt分布



標本平均の分布３.

Theorem (検定の根拠となる)

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。

展開してみる

$$t = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{\frac{(n-1)S/\sigma^2}{n-1}}}$$

前述する中心極限定理により$\sqrt{n}(\bar{X}-\mu)/\sigma$が正規分布、標本分散の分布に関する結果により$(n-1)S/\sigma^2$が自由度$n-1$の$\chi^2$分布。従って、$t$が自由度$n-1$の t 分布に従う

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。

基本的な考え方

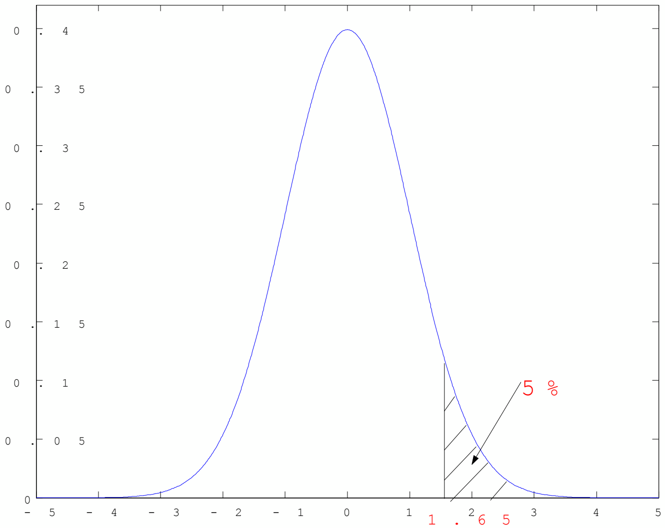
- まずXは平均（期待値）が$\mu \leq 20000$、標準偏差が$\sigma = s = 7500$の正規分布$N\left(20000, 7500^2\right)$に従うと仮定する。
- 上述の仮定の下で、$\bar{X}$が実現値20200を超える確率を調べる。
- $\bar{X}$が実現値20200を超える確率は極めて小さいなら、仮定が$\bar{X}$の実現値20200、すなわちデータと矛盾することを意味する。$\mu \leq 20000$の仮説が偽であると判断する（棄却する）。
- $\bar{X}$が実現値20200を超える確率は小さくない場合、$\mu \leq 20000$の仮説が真であると判断する（採択する）。

有意水準　確率が小さいかどうかを判断する基準である。通常αで表す、慣例として1%か5%とする。

臨界値（有意水準点）　確率が有意水準aに対応している確率変数の値。たとえば、今の例では$P(X > c) = a$となるようなcの値。

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、$P(X > 1.65) = 5\%$。



解答

- 有意水準を $\alpha = 5\%$とする。
- 帰無仮説 H_0：期待値 $\mu \le 20000$； 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 $\mu \le 20000$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

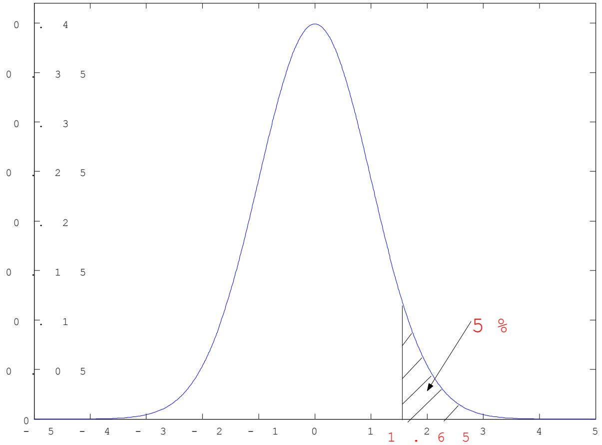
注

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臨界値と比較する。t 値が左側の臨界値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. 左側の臨界値は分布表から見つかった右側の臨界値の値 $\times(-1)$)、たとえば、右側の 5% 臨界値が 1.65 だったら、左側は -1.65 となる。

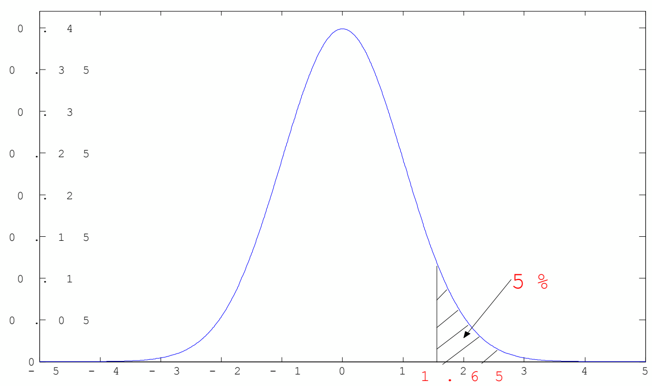
密度関数のグラフで見る

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臨界値よりも右にあって、$\mu \leq 20000$ の仮説が偽であると判断し棄却する。

棄却域 影の部分は棄却域と呼ぶ。検定統計量 t 値が棄却域に入った場合、帰無仮説は棄却される。



1. 対立仮説はより大きい（>）となっている場合、棄却域はグラフの右裾にある。t値が棄却域に入った場合（すなわちは検定統計量tが臨界値よりも右にある場合）、帰無仮説を棄却する。
2. 対立仮説はより小さい（<）となっている場合、棄却域はグラフの左裾にある。t値が棄却域に入った場合（すなわちは検定統計量tが臨界値よりも左にある場合）、帰無仮説を棄却する。
3. 以下の図は対立仮説はより大きい（>）となっている場合の棄却域を示している。



解答のステップ

1. 有意水準を選ぶ。例の場合、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。例の場合、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%臨界値の値を読み取る。例の場合、臨界値が約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。例の場合、$t = 100 \times (20200 - 20000)/7500 = 2.66$。
5. ステップ3で求めたtとステップ2で求めた臨界値の値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。例の場合、$t > 1.65$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：$\mathrm{H}_0 : \mu \le 10$；$\mathrm{H}_1 : \mu > 10$。

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 $\bar{X} = 3.2\,\mathrm{cm}, s = 2$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、$\mathrm{H}_0 : \mu = 3$；$\mathrm{H}_1 : \mu > 3$または$\mu < 3$。
2. 有意水準を決めてから、それを半分にして、臨界値の値を調べる。たとえば、有意水準を 5%にした場合、それの半分2.5%の臨界値の値を調べる。
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の絶対値と臨界値の値（たとえば、2.5%の臨界値の値）と比較する。$|t| >$臨界値の値であれば、帰無仮説を棄却する、逆に$|t| \le$臨界値の値であれば、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。
- $\mathrm{H}_0: \mu = 3$; $\mathrm{H}_1: \mu > 3$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% 臨界値が 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。

グラフで見る両側検定

- 両側検定の棄却域は両裾の影の部分に対応する。t値がどちらに入っても帰無仮説H_0は棄却される。