

統計学

北門 利英（海洋生物資源学科）

Lecture 8 正規分布における信頼区間の構成

Attention:

- 大学から「前学期の定期試験は原則としては対面形式での実施は行わないこと」との指示がありましたので9月に予定していた対面試験を行わず、代替方式を今後検討し後日連絡致します。
- これに安心せず、今回およびこれまでの授業に関してわからないことがあれば、メールカリアルタイム接続時に遠慮なく質問してください（7月1日も13:00-14:00とします）。

Point: 前回は正規分布 $N(\mu, \sigma^2)$ に対して、 μ の値がいくらかを推定する点推定について学んだ。今回は、 μ の値がどの程度の範囲にあると考えられるかを問う区間推定について学ぶ。次の用語をしっかりと理解すること。

- 分散が既知の場合に、正規分布の平均値に対する信頼区間が求められること
- 分散が未知の場合に、正規分布の平均値に対する信頼区間が求められること

T. Kitakado

Statistics

Lecture 08, July 1, 2020

生物資源学科の北門です。このビデオは統計学のLecture8の要約版です。今回は、**正規分布における信頼区間の構成**についてお話します。授業の前に連絡ですが、当初9月の定期試験期間に対面試験が可能な状況となっていれば対面で試験を行うとしましたが、先週の大学からの連絡では原則として対面形式で試験を行わないこと、と連絡がありましたので、別途課題を出すなどの代替の方法を考えてます。これについては後日連絡いたします。それでは始めます。

正規分布とパラメータの点推定の復習

T. Kitakado

Statistics

Lecture 08, July 1, 2020

最初に再度正規分布の復習です。

正規分布の定義

連続型分布 1 確率変数 Y の確率密度関数が

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty \quad \left(\begin{array}{l} -\infty < \mu < \infty \\ 0 < \sigma < \infty \end{array} \right)$$

となるとき, Y は正規分布 (normal distribution) $N(\mu, \sigma^2)$ にしたがうという。

期待値
(平均値)
分散

$$\int_{-\infty}^{\infty} f(y) dy = 1$$

$$E[Y] = \int_{-\infty}^{\infty} y f(y) dy = \mu$$

$$V[Y] = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy = \sigma^2$$

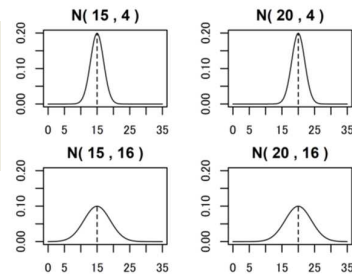


図4 正規分布の確率密度関数の概形

T. Kitakado

Statistics

Lecture 08, July 1, 2020

正規分布は連続型の確率分布で, 確率密度関数がこの式のように表される場合でした。muは期待値, sigma2は分散を意味しています。また確率密度関数の概形は平均値に対して左右対称で, 分散sigma2が大きくなると図でもわかりますように平均値の周りの散らばりが大きくなります。

正規分布に関する性質 (1) 標準正規分布 $N(0,1)$

性質 1 確率変数 Y が $N(\mu, \sigma^2)$ にしたがうとき,

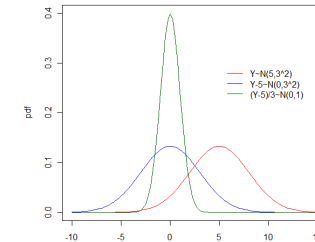
$$Z = \frac{Y - \mu}{\sigma} \quad (5)$$

は $N(0,1)$ にしたがう。この変換を標準化 (standardization), また $N(0,1)$ を標準正規分布 (standard normal distribution) という。

$$Y \sim N(\mu, \sigma^2)$$

$$Y - \mu \sim N(0, \sigma^2)$$

$$Z = \frac{Y - \mu}{\sigma} \sim N(0,1)$$



T. Kitakado

Statistics

Lecture 08, July 1, 2020

正規分布に関係する大切な性質として標準化という話をしました。これは, 確率変数 Y が平均 μ , 分散 σ^2 の正規分布に従っているとき, この Y から平均値 μ を引いて, 標準偏差 σ で割った値を Z と置くと, もとものの μ と σ が何であってもそれを使って変換すれば平均が 0, 分散が 1 の正規分布に変換できるということでした。このような変換を標準化, また平均 0, 分散 1 の正規分布を標準正規分布と呼びます。

定式化 (μ の推定)

$$Y_1, Y_2, \dots, Y_n \sim (iid) N(\mu, \sigma^2)$$

$$\hat{\mu}(Y) = \bar{Y} = \frac{1}{n}(Y_1 + \dots + Y_n)$$

推定量の不偏性

$$E[\hat{\mu}(Y)] = E\left[\frac{1}{n}(Y_1 + \dots + Y_n)\right] = \frac{1}{n}(E[Y_1] + \dots + E[Y_n]) = \mu$$

推定量の分散

$$V[\hat{\mu}(Y)] = V\left[\frac{1}{n}(Y_1 + \dots + Y_n)\right] = \frac{1}{n^2}(V[Y_1] + \dots + V[Y_n]) = \frac{\sigma^2}{n}$$



T. Kitakado

Statistics

Lecture 08, July 1, 2020

さて正規分布の場合の平均値 μ の推定ですが、標本平均を用います。この推定量は、ここに示していますように不偏で、また証明は省略しますが分散を σ^2/n よりも小さくする不偏な推定量は存在しないことが示せます。

定式化(σ^2 の推定)

$$Y_1, Y_2, \dots, Y_n \sim (iid) N(\mu, \sigma^2)$$

$$\hat{\sigma}^2(Y) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

推定量の不偏性(不偏であることだけを覚えてください)

$$E[\hat{\sigma}^2(Y)] = E\left[\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

推定量の分散(これは覚えなくてよいです)

$$V[\hat{\sigma}^2(Y)] = \frac{2\sigma^4}{n-1}$$



T. Kitakado

Statistics

Lecture 08, July 1, 2020

また分散ですが、各観測値と標本平均の差の2乗を平均するわけですが、ここで n ではなく $(n-1)$ で割っているのは、その下の式に示したように、そうすることで不偏性を満たすことができるからです。

区間推定法(1)

正規分布の平均値 μ に対する信頼区間 (分散既知の場合)

T. Kitakado

Statistics

Lecture 08, July 1, 2020



最初に分散既知の場合の正規分布の平均値 μ に対する信頼区間についてです。

例題

例 1 (前回のマグロの続き) ある海域に生息するマグロ種の 1 歳魚を 9 個体ランダムにサンプリングし、その体長を測定したところ 42, 40, 50, 39, 40, 43, 42, 48, 43 (cm) であった。これらが独立同一に正規分布 $N(\mu, \sigma^2)$ にしたがうとする。ただし σ は既知で 6 とする。このとき、90% 信頼区間, 95% 信頼区間および 99% 信頼区間をそれぞれ求めよ。

$$Y_1, Y_2, \dots, Y_n \sim (iid) N(\mu, \sigma^2)$$

- これまでは、パラメータの値を「1点」で推定した
(例えば, $\hat{\mu} = 43\text{cm}(SE = 1.23\text{cm})$ などのように)
- これとは別に, 「区間」で推定する方法もある
(例えば, μ の 95% 信頼区間 = (39.08, 46.92) など)

T. Kitakado

Statistics

Lecture 08, July 1, 2020



ここでは前回のマグロの体長データを例にとります。これまでは、パラメータの値を $\mu\text{-hat}=43\text{cm}$ のように点で推定してきました。これを今回の区間推定と区別するために、点推定とも言います。今回は、この点推定とは別にパラメータの値を区間で推定することを考えたいと思います。

区間推定の定式化

1. 観測データの確率分布と推定するパラメータを定義
2. 区間推定の信頼水準を定める(例えば95%とすると)
この信頼水準の意味は、信頼区間区間が本当のパラメータの値を含む頻度が95%であることをいう
3. 信頼水準を満たすような区間の決め方を定める
4. 観測データを得たとき、3の方式に基づいて区間を計算し、これを信頼水準95%の信頼区間として与える

T. Kitakado

Statistics

Lecture 08, July 1, 2020



区間推定の定式化 (例)

1. 観測データの確率分布と推定するパラメータを定義

$$Y_1, Y_2, \dots, Y_n \sim (iid) N(\mu, \sigma^2)$$

2. 信頼水準を95%とする
3. 信頼水準を満たすような区間の決め方を定める

$$\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}$$

4. 観測データに対して計算

例えば、 $n=9, \sigma=6, \bar{Y} = 43$ のとき $36.08 \leq \mu \leq 43.92$

T. Kitakado

Statistics

Lecture 08, July 1, 2020

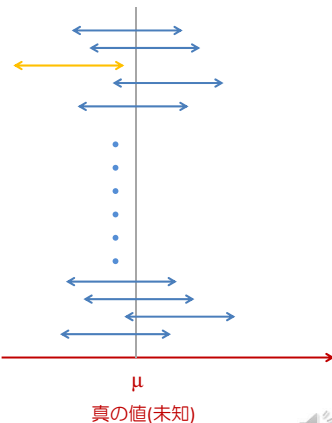


これを数式を用いてシンプルに表すとこのようになります。

一般的な区間推定の定式化ですが、まず観測データの確率分布と推定するパラメータを定義します。今回は、正規分布における平均値パラメータ μ が対象となります。次に、例えば95%など区間推定の信頼水準を定めます。この信頼水準の意味は、信頼区間が本当のパラメータの値を含む頻度が信頼水準と一致することをいいます。次のスライドでもう少し詳しく説明します。次に、その信頼水準を満たすような区間の決め方を定めます。そして、実際の観測値を得たら3で決めた方式に基づいて区間を計算し、これを信頼水準95%の信頼区間として与える

信頼水準の意味をもう少し

- 信頼水準を95%と設定する
- 観測値に対して信頼区間を計算する
- 観測値を得て信頼区間を計算するという操作を何度も繰り返したとすると(例えば100回), そのうち95%の頻度で真の μ の値を含む



T. Kitakado

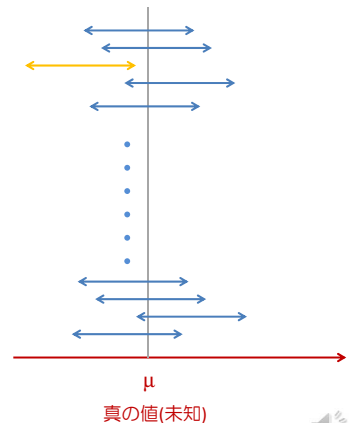
Statistics

Lecture 08, July 1, 2020

さて信頼水準の意味ですが, 例えば信頼水準を95%と設定します。そしてある方式に基づいて信頼区間を計算したとします。信頼水準の意味は, 観測値を得て信頼区間を計算するという操作を仮に何度も繰り返したとすると(例えば100回としましょう), そのうち95回すなわち95%の頻度で真の μ の値を含んでいることを意味します。

信頼水準の意味をもう少し

- 信頼水準を95%と設定する
- 観測値に対して信頼区間を計算する
- 観測値を得て信頼区間を計算するという操作を何度も繰り返したとすると(例えば100回), そのうち95%の頻度で真の μ の値を含む
- データの解析者は1セットしかデータをもっていないので, そのデータから構成した信頼区間が真の値を含んでいるどうかは分からない(確率的にしか分からない)



T. Kitakado

Statistics

Lecture 08, July 1, 2020

さて信頼水準の意味ですが, 例えば信頼水準を95%と設定します。そしてある方式に基づいて信頼区間を計算したとします。信頼水準の意味は, 観測値を得て信頼区間を計算するという操作を仮に何度も繰り返したとすると(例えば100回としましょう), そのうち95回すなわち95%の頻度で真の μ の値を含んでいることを意味します。

ところで信頼区間をどのように構成？

- 定めた水準 $1-\alpha$ に合うよう、観測データの確率分布に基づき区間を構成する (95%信頼水準なら $\alpha=0.05$)

$$Y_1, Y_2, \dots, Y_n \sim (iid) N(\mu, \sigma^2)$$

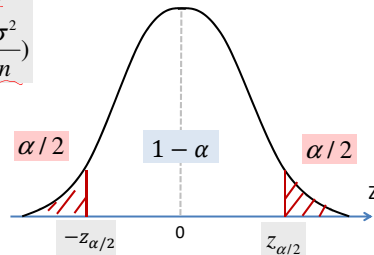
$$Y_1 + Y_2 + \dots + Y_n \sim N(n\mu, n\sigma^2)$$

$$\bar{Y} = \frac{Y_1 + Y_2 + \dots + Y_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{Y} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0, 1)$$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



T. Kitakado

Statistics

Lecture 08, July 1, 2020

さて正規分布における平均値パラメータの信頼区間の構成方法ですが、正規分布の性質として、 $\bar{Y}$ の標本平均の確率分布が...のように求まりますので、 Z を...のように変形すればそれが標準正規分布に従うことが分かります。これは未知である μ の値が何であっても成り立ちます。ということは、水準が $1-\alpha$ なら、この範囲がその確率となります。

ところで信頼区間をどのように構成？

- 定めた水準 $1-\alpha$ に合うよう、観測データの確率分布に基づき区間を構成する (95%信頼水準なら $\alpha=0.05$)

$$Y_1, Y_2, \dots, Y_n \sim (iid) N(\mu, \sigma^2)$$

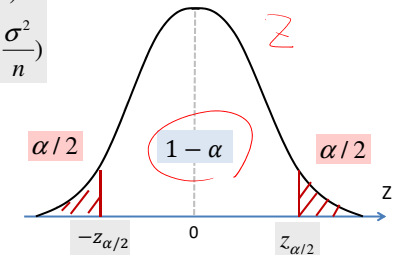
$$Y_1 + Y_2 + \dots + Y_n \sim N(n\mu, n\sigma^2)$$

$$\bar{Y} = \frac{Y_1 + Y_2 + \dots + Y_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\bar{Y} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0, 1)$$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



T. Kitakado

Statistics

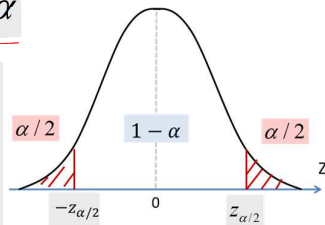
Lecture 08, July 1, 2020

さて正規分布における平均値パラメータの信頼区間の構成方法ですが、正規分布の性質として、 $\bar{Y}$ の標本平均の確率分布が...のように求まりますので、 Z を...のように変形すればそれが標準正規分布に従うことが分かります。これは未知である μ の値が何であっても成り立ちます。ということは、水準が $1-\alpha$ なら、この範囲がその確率となります。

ところで信頼区間をどのように構成？

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

$$\begin{aligned} & -z_{\alpha/2} \leq Z \leq z_{\alpha/2} \\ \Leftrightarrow & -z_{\alpha/2} \leq \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} \leq z_{\alpha/2} \\ \Leftrightarrow & -z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{Y} - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow & -\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow & \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow & \bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \end{aligned}$$



T. Kitakado

Statistics

Lecture 08, July 1, 2020

すなわち、区間をこのように変形して μ に関する不等式にすると、このように信頼水準 $1-\alpha$ の μ の信頼区間が求まります。

演習(1)

先程のマグロの体長の例で9個体のデータが下記のとおりであった。 $\sigma=6$ と仮定し、90%信頼区間, 95%信頼区間および99%信頼区間を求めよ。

$Y=(42, 40, 50, 39, 40, 43, 42, 48, 43)$

90%信頼区間の場合

$$\begin{aligned} & \bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \Leftrightarrow & 43 - 1.645 \frac{6}{\sqrt{9}} \leq \mu \leq 43 + 1.645 \frac{6}{\sqrt{9}} \\ \Leftrightarrow & 39.71 \leq \mu \leq 46.29 \end{aligned}$$

T. Kitakado

Statistics

Lecture 08, July 1, 2020

具体的な例で計算してみると、...

演習(1)

先程のマグロの体長の例で9個体のデータが下記のとおりであった。 $\sigma=6$ と仮定し、90%信頼区間, 95%信頼区間および99%信頼区間を求めよ。

$Y=(42, 40, 50, 39, 40, 43, 42, 48, 43)$

90%信頼区間 = (39.71, 46.29) $z=1.645$

95%信頼区間 = (39.08, 46.92) $z=1.960$

99%信頼区間 = (37.85, 48.15) $z=2.575$



T. Kitakado

Statistics

Lecture 08, July 1, 2020

他の信頼水準にたいしても同様に計算が可能です。ここで理解してほしいのは、信頼水準の意味と信頼区間の計算の方法ですが、水準を大きくすると、それだけ真の μ を含まないという失敗が許されませんから、自ずと区間が広くなることが分かります。

区間推定法(2)

正規分布の平均値 μ に対する信頼区間
(分散未知の場合)



T. Kitakado

Statistics

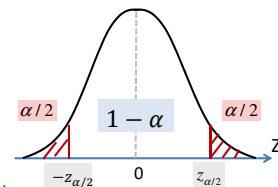
Lecture 08, July 1, 2020

次に分散が未知の場合の正規分布の平均値 μ に対する信頼区間についてです。

分散が既知の場合と未知の場合の違いは？

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



残念ながら、いま上記の式は使えない
(なぜなら、分散が未知だから！)
そこで、分散の推定値を代入するが、正規分布の仮定が崩れる

T. Kitakado

Statistics

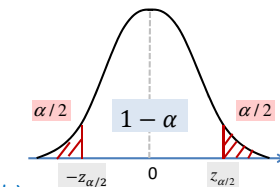
Lecture 07, June 24, 2020

先ほどは分散 σ^2 が既知でしたので、このように標準化することができましたが、分散が未知の場合には推定しなければなりません。推定の方法は前回説明しましたが、このような式で求めることができます。ところが、この推定した σ の値を代入して標準化しても、真の σ を代入したわけではなく、不確実な σ を代入しているので標準化した値は標準正規分布とはならず、この授業で初めて出てきますが自由度 $n-1$ のt分布と呼ばれる確率分布に従います。ただあまり難しく考えないでください。標準正規分布と形は同じで、少しこのような裾の値が異なる点を理解してください。また自由度がt分布の性質を規定しますが、データの数によって変わると理解してください。

分散が既知の場合と未知の場合の違いは？

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



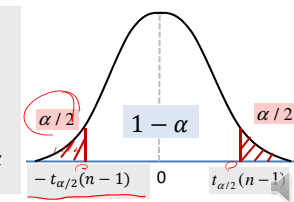
残念ながら、いま上記の式は使えない
(なぜなら、分散が未知だから！)
そこで、分散の推定値を代入するが、正規分布の仮定が崩れる

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$T = \frac{\bar{Y} - \mu}{\hat{\sigma} / \sqrt{n}} \sim t(n-1)$$

自由度 $n-1$ のt分布

$$P(-t_{\alpha/2}(n-1) \leq T \leq t_{\alpha/2}(n-1)) = 1 - \alpha$$



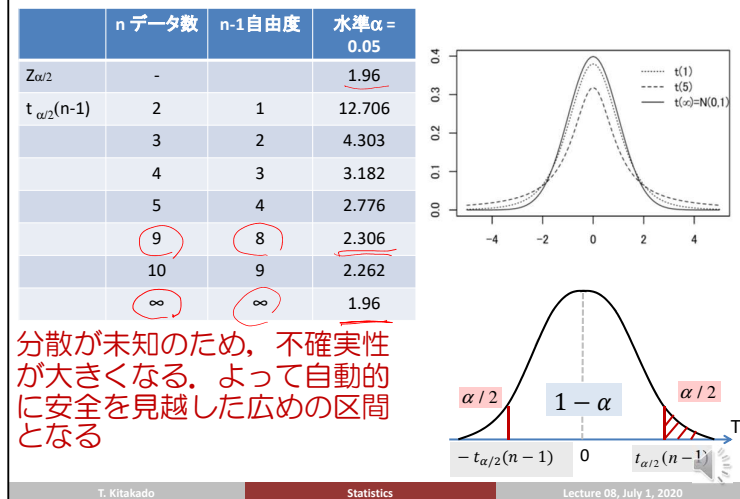
T. Kitakado

Statistics

Lecture 08, July 1, 2020

先ほどは分散 σ^2 が既知でしたので、このように標準化することができましたが、分散が未知の場合には推定しなければなりません。推定の方法は前回説明しましたが、このような式で求めることができます。ところが、この推定した σ の値を代入して標準化しても、真の σ を代入したわけではなく、不確実な σ を代入しているので標準化した値は標準正規分布とはならず、この授業で初めて出てきますが自由度 $n-1$ のt分布と呼ばれる確率分布に従います。ただあまり難しく考えないでください。標準正規分布と形は同じで、少しこのような裾の値が異なる点を理解してください。また自由度がt分布の性質を規定しますが、データの数によって変わると理解してください。

分散が既知の場合と未知の場合の違いは？



つまり、sigmaが既知であれば標準正規分布の性質が使えますので、たとえば95%信頼区間であれば、95%の確率で-1.96から1.96の範囲にZが入る、という性質が使えました。一方で分散が未知の場合、データの数に応じてt分布の形も変わり、またそれぞれの水準にあった区間も変わってきます。ただ、データの数が多くなると分散の推定精度が上がり、データ数が無限大の場合には分散既知の場合と同等となることも分かります。

分散未知の場合の信頼区間

$$T = \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-1)$$

$$P(-t_{\alpha/2}(n-1) \leq T \leq t_{\alpha/2}(n-1)) = 1 - \alpha$$

$$-t_{\alpha/2}(n-1) \leq \sqrt{n} \frac{\bar{Y} - \mu}{\hat{\sigma}} \leq t_{\alpha/2}(n-1) \quad \text{分散未知の場合}$$

$$\Leftrightarrow \bar{Y} - t_{\alpha/2}(n-1) \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\alpha/2}(n-1) \frac{\hat{\sigma}}{\sqrt{n}}$$

$$\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{分散既知の場合}$$

T. Kitakado Statistics Lecture 08, July 1, 2020

このt分布の性質を用いて、最終的にはmuの信頼区間がこうのように求まります。分散が既知と未知との違いは、分散の推定値を代入するかどうかと、標準正規分布を仮定すべきか、t分布を仮定するべきか、の2点です。ここをしっかりと区別してください。

演習(2)

先程のマグロの体長の例で9個体のデータが下記のとおりであった。σの推定値は3.71と求まる。このとき、90%信頼区間、95%信頼区間および99%信頼区間を求めよ。

90%信頼区間の場合



$$\begin{aligned} \bar{Y} - t_{\alpha/2}(n-1) \frac{\hat{\sigma}}{\sqrt{n}} &\leq \mu \leq \bar{Y} + t_{\alpha/2}(n-1) \frac{\hat{\sigma}}{\sqrt{n}} \\ \Leftrightarrow 43 - 1.860 \frac{3.71}{\sqrt{9}} &\leq \mu \leq 43 + 1.860 \frac{3.71}{\sqrt{9}} \\ \Leftrightarrow 40.70 &\leq \mu \leq 45.30 \end{aligned}$$

T. Kitakado

Statistics

Lecture 08, July 1, 2020

分散が未知の場合の信頼区間の計算方法です。

演習(2)

先程のマグロの体長の例で9個体のデータが下記のとおりであった。σの推定値は3.71と求まる。このとき、90%信頼区間、95%信頼区間および99%信頼区間を求めよ。

$$90\% \text{信頼区間} = (40.70, 45.30) \quad t=1.860$$

$$95\% \text{信頼区間} = (40.15, 45.85) \quad t=2.306$$

$$99\% \text{信頼区間} = (38.85, 47.15) \quad t=3.355$$

T. Kitakado

Statistics

Lecture 08, July 1, 2020

90%信頼区間、95%信頼区間および99%信頼区間ですが、ここでも水準を上げると区間が広がることが分かります。

提出課題

今回の Lecture 8 と次回の Lecture 9 の課題と併せて一つの提出課題とします。計算式と答えを書いて、必ずまとめて 1 つのファイルで提出してください。ワードで数式の書けない人は、手書きの計算結果を写真に撮ってワードに張り付けて下さっても構いません。答だけの回答は大きく減点します。提出期限は 7 月 17 日 23:59 とします。

Lecture08-HW1 ある生物種の 6 個体を同じ環境で飼育したところ、生存日数が 1472, 1486, 1401, 1350, 1610, 1585 日であった。この観測値が正規分布にしたがうと仮定し、平均値 μ の 90% 信頼区間, 95% 信頼区間および 99% 信頼区間をそれぞれ求めよ。ただし分散は未知とする。



T. Kitakado

Statistics

Lecture 08, July 1, 2020

最後に提出課題です。次回Lecture9の宿題と併せて7月17日までに学務システムを通して提出してください。それではLecture8のダイジェストビデオを終わります。