2 代表値

1	Excelデータ分析の準備	ファイル > オプション > アドイン > 設定 を開き、以下にチェック ・ソルバー アドイン ・分析ツール	p38-39	01-05
2	記述統計	平均を求めるなどの、データの傾向を捉える手法全般のこと	p42-44	02-01
3	平均値	データの重心。すべてのデータの合計をデータ数で割る	p46-48	02-02
4	平均値を求める関数	AVERAGE(範囲)	p46-48	02-02
5	中央値	データを昇順または降順に並べたときに、真ん中にくる値。 データ数が偶数の場合は、真ん中の二つの平均値。	p49-50	02-03
6	中央値を求める関数	MEDIAN(範囲)	p49-50	02-03
7	最頻値	データ群の中で、最も多く出現する値		
8	最大値・最小値	データ群の中で、最も大きい値・または最も小さい値。 それぞれ以下の関数で求められる。 最大値：MAX(範囲) 最小値：MIN(範囲)	p55	02-05
9	外れ値	他のデータと比べて、極端に大きい、または小さい値。 平均値は外れ値の影響を受けやすく、中央値・最頻値は外れ値の影響を受けにくい。	p55	02-05
10	ヒストグラム	データの傾向を読み取るために、データ分布の形状を可視化したチャート	p80-88	03-02

授業コ：データ分布の形状によって、データ群の代表値として適切なものが平均値か、中央値か、頻出値か、変わってきます。まずはデータの傾向をとらえ、

3 分散と標準偏差

1	分散	平均値を中心に、データがどの程度ばらついているかを表す	p51-52	02-04
2	偏差	平均値との差。 データの値 - 平均値	p51-52	02-04
3	分散の求め方	①各データの偏差を求める ②偏差を二乗する（偏差平方） ③を合計する（偏差平方和） ④③をデータ数-1で割る	p51-52	02-04
4	分散の求め方の①で二乗する理	偏差をすべて合計すると＋と－が打ち消し合って0に近い値になってしまうので	p51-52	02-04
5	分散を求める関数	VAR.S(範囲)	p51-52	02-04
6	標準偏差	分散の平方根。略してSD（Standard Deviation）。 分散を求める過程で二乗しているので、元に戻す操作をしている。 Excelで平方根を求めるには、SQRT(値) 関数を使う。	p54-55	02-05
7	標準偏差を求める関数	STDEV.S(範囲)	p54-55	02-05
8	標準化	平均=0、標準偏差=1に合わせて整形する方法。 各データの偏差を標準偏差で割る。		
9	標準得点	標準化された値		
10	偏差値	標準得点*10 + 50 全体の中での相対的な位置を知るために用いられる		

授業コ：統計を理解するうえで、分散と標準偏差の理解はもっとも基本となります。計算過程が何を行っているものなのか理解しましょう。

4 相関

1	散布図	２つの量的変数（連続変数）の関係を見るためのチャート	p98-100	03-05
2	線形近似	散布図にプロットされたデータの傾向をなぞる線を引くことを線形近似という。 データが線のように分布している場合、線形近似が有効。 近似とは細部を無視して、単純化すること。	p101-105	03-05
3	回帰式	線形近似で引かれた線を表す方程式	p101-105	03-05
4	相関関係	片方の変数がもう片方の変数に一定の影響を与えあう関係のこと。 線形に関係する場合のみを対象とする。	p109-115	03-06
5	正の相関と負の相関	正の相関：変数Ｘの増減によって、変数Ｙも比例して増減する関係 負の相関：変数Ｘの増減によって、変数Ｙが反比例して増減する関係	p109-115	03-06
6	共分散	Σ((X - Xの平均)*(Y - Yの平均)) / (データ数-1) ① 変数Ｘの各データ - 変数Ｘの平均 を求める ② 変数Ｙの各データ - 変数ＹBの平均 を求める ③ ①と②を掛け合わせる ④ ③の合計を求める ⑤ ④をデータ数-1で割る	p109-115	03-06
7	相関係数	共分散 / (変数Ｘの標準偏差 * 変数Ｙの標準偏差) 1または-1に近いほど相関が高く、0に近いほど相関が低い。 ±0.5付近なら弱い相関あり、±0.7を超えると相関が高い、という傾向がある。	p109-115	03-06
8	相関行列	複数の変数がある場合に、その中の２つの変数同士の相関係数を羅列した行列	p116-120	03-07
9	因果関係	原因と結果の関係	p108	03-05
10	疑似相関	相関関係はあるが、因果関係がない、という関係のこと。 ２つの変数に共通する別の変数の存在がある場合や、たまたまデータの傾向が似てしまった場合などが挙げられる。	p108	03-05

授業コ：２変数間の関係を調べる場合、散布図を描くこと、相関係数を見ること、疑似相関でないか確認すること、の３点は必ず行いましょう。

5 確率分布

1	母集団	すべてのデータ。センサーから温度を取得するIoTシステムであれば、すべての	p130	04-04

2	標本	母集団から抜き出したデータ。抜き出す作業をサンプリングという。 すべてのデータを対象に分析するのは現実的でないため、標本を抜き出す。 IoTシステムにおいては、1分おきに取得した1週間分のセンサーデータ、など。	p130	04-04
3	確率変数	確率に対応する値。サイコロのそれぞれの目が出る確率はいずれも1/6であるが、	p132-134	04-04
4	確率分布	横軸に確率変数、縦軸に確率をとるグラフ。 ヒストグラムの幅を限りなく細くしたもの。 全体の面積が1となる。	p132-134	04-04
5	一様分布	区間内のどの値をとる確率も等しい確率分布。 サイコロを振ってそれぞれの目が出る確率など。		
6	二項分布	二者択一の試行を複数回繰り返したときの確率分布。	p132-134	04-04
7	正規分布	左右対称の確率分布。平均値、中央値、最頻値がすべて中央となる。	p132-134	04-04
8	中心極限定理	平均値をたくさん取り出してプロットすると、必ず正規分布となる	p136-137	04-05
9	正規分布と標準偏差の関係	±1σの範囲に約68.3%、±2σの範囲に約95.4%、±3σの範囲に約99.7%が含まれる ※σ（シグマ）とは、標準偏差のこと。	基本情報技術者テキストp46	
10	標準正規分布	平均や標準偏差が異なると正規分布の形状が異なるので、計算しやすくするため	p135,149-150	04-04

授業コメ：次回以降は、データが正規分布に従っていることを前提として、データの分析を行っていきます。

6 区間推定

1	推測統計	標本を元に、母集団のデータの特徴を推測すること。
2	母集団と標本	母集団：すべてのデータ 標本：母集団から抜き出した一部のデータ
3	母平均と標本平均	母平均：母集団の平均値。 標本平均：標本の平均値。
4	点推定	母平均を推定する方法の1つで、1点に絞った推定。 たとえば、「標本平均が50であるとき、母平均も50である」とする考え方。
5	区間推定	母平均を推定する方法の1つで、区間（範囲）で絞った推定。 たとえば、「母平均は30～70の間である」など。
6	信頼度	区間推定において、推定した範囲内に母平均が入る可能性を示す値。 多くの場合、95%を利用する。
7	信頼区間	点推定値を中心とした区間。95%信頼区間が一般的に利用される。 95%信頼区間とは、確率分布のうち、95%分の面積となる範囲。 標本を抽出して95%信頼区間を求める作業を100回行ったときに、95回は母平均がその区間内に含まれるという意味。
8	信頼区間の求め方	正規分布では、±1σの範囲に68.3%、±2σの範囲に95.4%、±3σの範囲に99.7%が含まれるので、 95%信頼区間は、±1.96σに挟まれた範囲となる。
9	標準誤差	標準偏差 / √データ数 データ数が多いほど、標準誤差は小さくなる。 標準誤差が小さい＝標本平均が信用できる（誤差が少ない）
10	区間推定の方法	標本平均-(1.96*標準誤差) ～ 標本平均+(1.96*標準誤差) の範囲に母平均が入るであろうと考える

授業コメ：±1σ＝68.3%、±2σ＝95.4%、±3σ＝99.7%に加えて、±1.96σ＝95.0%も重要な数字です。この数字はそのまま暗記してしまいましょう。

7 検定の手順

1	仮説検定	差異が偶然なのか、そうでないのかを結論づけるために使われる手法。 偶然ではないことを「統計的に有意な差がある」という。	p122-125	04-01
2	仮説の立て方	①主張したいこと（対立仮説）とその反対の仮説（帰無仮説）を立てる ②帰無仮説が真であった場合に、標本データが得られる確率を計算する ③②の確率が低ければ（多くの場合、5%未満であれば）、帰無仮説は真ではないだろうと考え、対立仮説を採用する。	p126-129	04-01
3	帰無仮説	棄却されるべき仮説（通常、仮説を立てた人が正しくないと思っている仮説） 差異が偶然である（統計的な有意差がない）ことを主張する。 たとえば「平均値＝70」といったように、具体的な数値や状態を指した形式になっている必要がある。	p126-129, p138-1	04-06
4	対立仮説	採用されるべき仮説（通常、仮説を立てた人が正しいと思っている仮説） 差異が偶然でない（統計的な有意差がある）ことを主張する。	p126-129, p138-1	04-06
5	棄却域	確率分布のうち、起こりづらいと考えられる範囲	p142-144	04-06
6	有意水準	棄却域の面積。確率分布の面積なので、単位は%となる。 5%が利用されることが多いが、10%, 1%, 0.1%などが利用されることもある。	p142-144	04-06
7	t値	帰無仮説の確率分布上に標本データをプロットしたときの、横軸の値。t値以上	p142-144	04-06
8	p値	帰無仮説が真であった場合に、標本データ以上に極端な値が得られる確率。 有意水準5%の場合、p値が5%よりも小さければ帰無仮説を棄却する。	p142-144	04-06
9	第一種の過誤と第二種の過誤	第一種の過誤：帰無仮説が正しいのに棄却してしまうこと 第二種の過誤：帰無仮説が間違っているのに棄却しないこと	p144-146	04-06
10	片側検定と両側検定	確率分布における下側（左端）または上側（右側）のどちらかを棄却域とするのが片側検定、両方を棄却域とするのが両側検定。 両側検定の場合、下側と上側の面積の合計が有意水準となる。（有意水準5%であれば、下側2.5%、上側2.5%）		

授業コメ：仮説検定における基本の流れを理解しましょう。次回からは具体的な検定を行っていきます。

8 t検定①　～1標本のt検定～

1	t検定	主に、平均値に差があるかどうかを調べる目的で使われる検定。 ここでは、分析者が仮定したある値と、1つの標本から推測される母平均が異なるかどうかを検定する。	p122-129	04-01

2	標準正規分布	平均＝0、標準偏差＝1となるように標準化された正規分布。 標準化のための計算過程で母集団の標準偏差を利用するが、現実的に母集団の標準偏差を知ることは難しい。	p150-152	04-07
3	t分布	標本の標準偏差を使って標準化した確率分布を、t分布という。 標準正規分布とほぼ同じ形状をしているが、データ数が少ないほど標準正規分布から離れた形になる。 t検定は、このt分布を利用した検定である。	p150-152	04-07
4	1標本t検定の帰無仮説と対立仮説	帰無仮説：ある値と母平均に有意差はない（ある値＝母平均） 対立仮説：ある値と母平均に有意差はある（ある値 ≠ 母平均）	p128	04-03
5	t値の求め方	(標本平均 - 母平均) / (標本の標準偏差 / √データ数)	p153	04-07
6	自由度	データ数 － 1	p159-160	04-08
7	T.DIST(t値, 自由度, TRUE)	片側検定の場合に利用。 左端～t値までの累積確率（面積）を求める。 右端が棄却域に設定されている場合は、この結果を1から引くことで、p値が求められる。	p155-163	04-08
8	T.DIST.RT(t値, 自由度)	片側検定の場合に利用。 t値～右端までの累積確率（面積）を求める。		
9	T.DIST.2T(t値, 自由度)	両側検定の場合に利用。 左端 ～ －t値と、 t値 ～ 右端の累積確率（面積）の合計を求める。		
10	t検定の結論を出す	有意水準＞p値：帰無仮説を棄却できる（有意差ありと見る） 有意水準＜p値：帰無仮説は棄却できない（有意差なしと見る）	p161	04-08

授業コメ 片側検定と両側検定のどちらを利用すべきかは、何を知りたいかによって変わってきます。データの特性を元に、どのような仮説を立てるかが非常に

9 t検定②　～2標本のt検定～

1	対応のないt検定	異なる対象からサンプリングした2標本の平均値に差があるかどうかを調べる。	p164-169	04-09
2	対応のない2標本とは	クラスごとのテストの点数の平均値に差があるかどうかを調べる場合、 1組と2組の異なる対象から標本を得るため、「対応なし」となる。	p164-169	04-09
3	対応のあるt検定	同じ対象からサンプリングした2標本の平均値に差があるかどうかを調べる。		
4	対応のある2標本とは	同じクラスに対して、補習を行う前と後でテストの点数の平均値に差があるかどうかを調べる場合、 同じ学生を対象として標本を得るため、「対応あり」となる。 対応のあるt検定は、2標本のデータ数が同じである必要がある。		
5	Excelを用いた仮説検定	[データ] > [データ分析] からさまざまな検定を選択できる。	p164-169	04-09
6	Excelによる対応のないt検定	[t検定: 分散が等しくないと仮定した2標本による検定]	p164-169	04-09
7	Excelによる対応のあるt検定	[t検定: 一対の標本による平均の検定]		
8	2標本のt検定の帰無仮説	2つの標本の平均値に差はない（差は0である）	p164-169	04-09
9	2標本のt検定の対立仮説	2つの標本の平均値に差がある（差は0ではない）	p164-169	04-09
10	有意水準	Excelでは「α」で表記される	p164-169	04-09

授業コメ 対応あり・なしの違いを理解しましょう。対象となるデータによって、行うべき検定が変わる点に注意してください。

10 χ2検定（カイ二乗検定）

		A/Bテスト： 2種類の広告を出し、どちらが良い成果を得られるかテストする		
1	マーケティング用語	コンバージョン率（CVR）： 成果達成数 / 成果達成に至る工程のうち、最初の段階の人数 たとえば、ショッピングサイトの広告であれば、広告をクリックした人数のうち、何人が商品の購入にまで至ったかを表す割合。	p170-176	04-10
2	χ2検定とは	主に、独立性の検定などに用いられる。 2変数間に関連があるかどうかを調べたい場合などに利用。	p170-176	04-10
3	クロス表	2変数を行・列に配置する	p170-176	04-10
4	実測値	実際の値のこと。観測度数ともいう。	p170-176	04-10
5	期待度数	理論上の値のこと。 各行・各列の合計値を元に、割合に応じて分配した値をクロス表のセルに埋める。	p170-176	04-10
6	χ2検定の計算方法	実測値と期待度数の差を求めて、χ2分布という確率分布に当てはめる。	p170-176	04-10
7	Excelによるχ2検定	CHISQ.TEST(実測値範囲, 期待値範囲)	p170-176	04-10
8	χ2検定の仮説の設定	帰無仮説：2つの変数は無関係である（実測値と期待度数の間に差はない） 対立仮説：2つの変数には関係がある（実測値と期待度数の間に差はある）	p170-176	04-10
9	p値から結論を出す	有意水準＞p値：帰無仮説を棄却できる（2変数に関係あり） 有意水準＜p値：帰無仮説は棄却できない（2変数に関係なし）	p170-176	04-10
10	χ2検定の注意点	p値が小さいほど関係が強いという意味ではないので、注意すること。 p値の大小は、データ数に依存するところが大きい。 2変数の関係の強さは、相関係数などから知ることができる。		

授業コメ χ2検定も、t検定と並んでよく使われる検定手法です。検定手法に応じて帰無仮説・対立仮説の立て方が異なる点に注意しましょう。

11 前処理

1	前処理	データ分析を行うにあたって、データの不備を修正したり、分析に適した形にデータを整形すること		
2	欠損値	本来入力されているべきであるにも関わらず、未入力状態のデータ	p178-188	05-01
3	質的変数の欠損値の処理	A. 欠損値がある行ごと削除 B. 出現頻度がもっとも高い値で置き換える C. 「欠損あり」などの特別な値で置き換える	p178-188	05-01
4	量的変数の欠損値の処理	A. 欠損値がある行ごと削除 B. 0で置き換える C. 平均値で置き換える D. 中央値で置き換える	p178-188	05-01

5	表記ゆれ	同じ意味の言葉が、微妙に異なる表現をされている状態のこと	p189-191	05-02
6	表記ゆれの処理方法	同じ意味の言葉の表現方法を統一して、表記ゆれが起きている値を置き換える	p189-191	05-02
7	外れ値・異常値	外れ値：データの中で、その他の多数のデータから大きく離れた値を持つもの 異常値：外れ値の中で、記入ミスや測定ミスなど、外れ値となった原因が何らかの間違いによるものだとわかっているもの	p192-197	05-03
8	外れ値・異常値の処理方法	次回扱う回帰分析では、外れ値は除外したほうが分析しやすい。 異常値は欠損値と同様の処理を行うことが多い。	p192-197	05-03
9	ダミー変数	質的変数（カテゴリカル変数）の値ごとに、0か1を割り当てる。 次回扱う回帰分析などを行うために必要な準備。 質的変数の値の種類が非常に多い場合は、それぞれに異なる数値を割り当てる「ラベルエンコーディング」と呼ばれる方法を取ることもある。	p198-204	05-04
10	ダミー変数の設定方法	①質的変数の値の種類と同じ数だけ、列を作る（回帰分析などを行う場合は、1列削除する） ②各列に、該当する場合は1、該当しない場合は0を設定する	p198-204	05-04

12 単回帰分析

1	回帰分析	原因となる変数と、結果となる変数の関係性を明らかにすることで、未来の予測などに利用することができる。 結果と原因の2要素をグラフにプロットしたとき、直線となる関係の回帰分析を線形回帰分析という。 また、原因となる変数が1つである場合の回帰分析を単回帰分析という。	p206-208	06-01
2	説明変数	原因となる変数のこと。独立変数ともいう。	p207　図6-1-1	06-01
3	目的変数（被説明変数）	結果となる変数のこと。従属変数ともいう。	p207　図6-1-1	06-01
4	最小二乗法	各データと回帰直線までの差の二乗合計が最小となるように切片と傾きを求める	p209　図6-2-1	06-02
5	モデル	モデルとは、データを解釈する方法を単純化したもの。 回帰分析におけるモデルは、説明変数と目的変数の関係を表す式（回帰式）のこと。	p209-211	06-02
6	回帰式	目的変数 ＝ 切片 + 傾き * 説明変数	p209-211	06-02
7	説明力	目的変数に対する、説明変数の影響の大きさ。 傾きが0であれば説明変数を変えても目的変数の値に影響を与えないため、傾きが0ではないことを「説明力がある」という。	p212-215	06-02
8	説明力があるかどうかの仮説検	帰無仮説：説明力がない（母集団において傾き＝0である） 対立仮説：説明力がある（母集団において傾き ≠ 0である）	p212-215	06-02
9	係数とp値	係数：傾き p値：説明変数に説明力があるかどうかの仮説検定の結果	p216-222	06-03
10	決定係数（R2 補正）	データに対して、回帰直線がどの程度フィットしているかを表す。 傾きのない横線を引いた場合の差の二乗合計と、回帰直線の差の二乗合計を比較したときに、何割減っているかを計算することで求められる。 0～1の範囲を取る値となり、だいたい0.5を超えてくると、そのモデルは悪くない精度とみなすことができる。	p216-222	06-03

授業コ：まずは回帰分析の概念や用語をおさえましょう。次コマではより実用的な重回帰分析を扱います。

13 重回帰分析

1	重回帰分析	説明変数が複数ある場合の回帰分析	p223-226	06-04
2	説明変数の説明力	説明変数が複数ある場合、それぞれの変数について説明力があるか（傾きが0で	p227	06-04
3	説明力に応じたモデル作成	p値が有意水準よりも大きい場合、その変数には説明力がないとみなし、モデルに含めない（回帰式から除外する） ただし、切片についてはp値が有意水準より大きくても、モデルに含めるのが一般的です。	p227	06-04
4	係数の解釈	説明変数を1増やしたときに、目的変数が係数分増えると解釈できる。 係数に大きすぎる数値が出た場合は、説明変数を0.1増やした場合に目的係数が係数*0.1分増える、といったように解釈する	p227	06-04
5	ダミー変数	ダミー変数は0か1なので、1であった場合のみその変数が目的変数に影響する、	p228-229	06-04
6	基準カテゴリ	ダミー変数を作るときに、削除した列。 基準カテゴリが1であった場合、他のダミー変数はすべて0であるはずなので、切片が削除した列の係数を示しているといえる。	p228-229	06-04
7	外れ値の処理	回帰分析を行う対象のデータに外れ値が含まれると、そちらに回帰直線がひっは	p231-234	06-05
8	多重共線性	説明変数同士に強い相関がある場合、それらの変数の傾きが不安定になってしまう現象。 相関がある変数は、どれか１つだけを説明変数に含めることで多重共線性を回避できる。	p234-238	06-05
9	重回帰分析における決定係数	重回帰分析を行うと、説明変数の数が多ければ多いほど「重決定 R2」は大きい値になってしまう傾向がある。 そこで、説明変数の数に合わせて調整した決定係数である「補正 R2」を使う。		
10	重回帰分析の流れ	①説明変数の相関行列を求める ②相関の高い変数を説明変数に2つ以上含めないようにしたうえで、回帰分析を実行 ③出力結果のp値を見て説明力がない変数の除外等をし、説明変数を絞って再度回帰分析を実行 ④決定係数を元に、ある程度精度の高いモデルであるかどうかチェック ⑤p値、決定係数ともに良い結果が得られれば、モデルを採用 ⑥回帰式の変数部分に任意の値（未来の値など）をあてはめて計算すると、ある説明変数を変更することで目的変数がどのように変化するのか、予測できる		