1. 記述統計

新しいデータを手に入れたときはいつでも，最初にやるべきことの一つは，データを簡単にまとめ，その傾向を理解しやすくする方法を見つけることです。これこそ記述統計の全てです (この反対は推測統計です)。実際，多くの人が “統計” という言葉を，記述統計の同義語だと思っています。この章で話そうとしているのがそれなのですが，詳細に入る前に，なぜ記述統計が必要なのかという感覚を掴んでもらいたいと思います。そうするためにまず，aflsmall_margins ファイルを開いて，ファイルの中にある変数を見てみましょう。

このように，一つの変数 afl.margins しかありません。この章ではこの変数に注目しますので，これが何なのか少し説明します。この本に含まれるデータセットとは違って，これは実際に得たデータであり，オーストラリアのフットボールリーグ (AFL) に関するデータです[*1]変数 afl.margins は，2010 年シーズンのホームゲーム，アウェイゲーム含めた全 176 ゲームの得点差 (獲得点数) が含まれています。

このアウトプットから，このデータが何を言おうとしているのか掴み取るのは簡単ではありません。“データを眺めている” だけでは，データを理解するのに全く効果的ではないのです。このデータが何を言おうとしているか，それを掴み取るためには，記述統計を計算する必要があり (この章で扱います)，わかりやすい図を描くことです（第 **??**章で扱います)。二つのやり方のうち，記述統計の方がより簡単なのですが，私たちが見ようとしているデータがどんなものなのかのイメージを掴むために，この afl.margins データのヒストグラムをお見せすることにしましょう。図 **??**を見てください。どうやってヒストグラムを描くかについては、セクション **??**で説明しますから。今は，ヒストグラムを見て それが afl.margins データを正しく理解する方法であることがわかってもらえれば結構です。

[*1] オーストラリア人ではない人にむけた注意：AFL はオーストラリアのルールで行われるフットボール競技です。この章を読むためにオーストラリアのルールを調べる必要は全くありません。

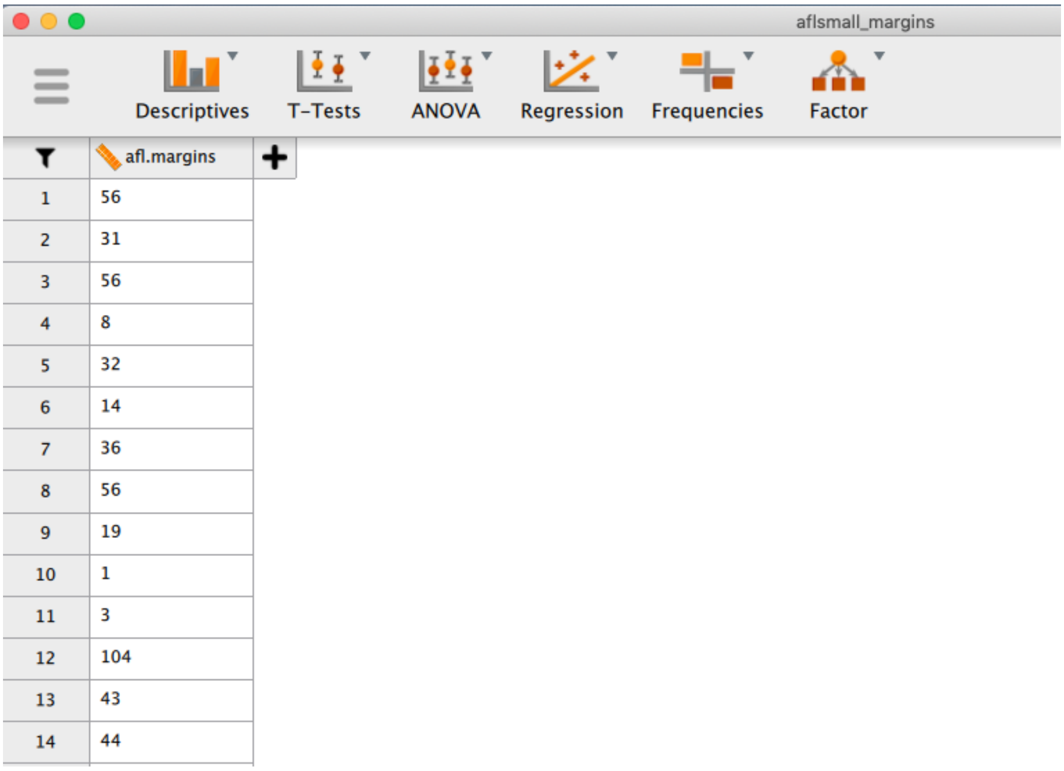


Figure1.1　JASP が `aflsmall_margins.csv` ファイルを開いて変数を見せているスクリーンショット

. . . .

1.1 傾向の測定

図 **??**で示したようなデータの絵を描くというのは，データがどうなっているのかの “要点” をもたらす優れた方法です。データをいくつかの単純な “集約された” 統計量に凝縮してみることが，特に便利です。いろんな場面で，まず計算してもらいたいのは**中心傾向**についての測定です。すなわち，あなたのデータの “平均’　’ や “真ん中’　’ がどのあたりにあるんのかを捉えて欲しいのです。最もよく使われる三つの数字は，平均値，中央値，最頻値です。これを順番に説明していきますので，その後でそれぞれがどういうときに便利なのかをみていきましょう。

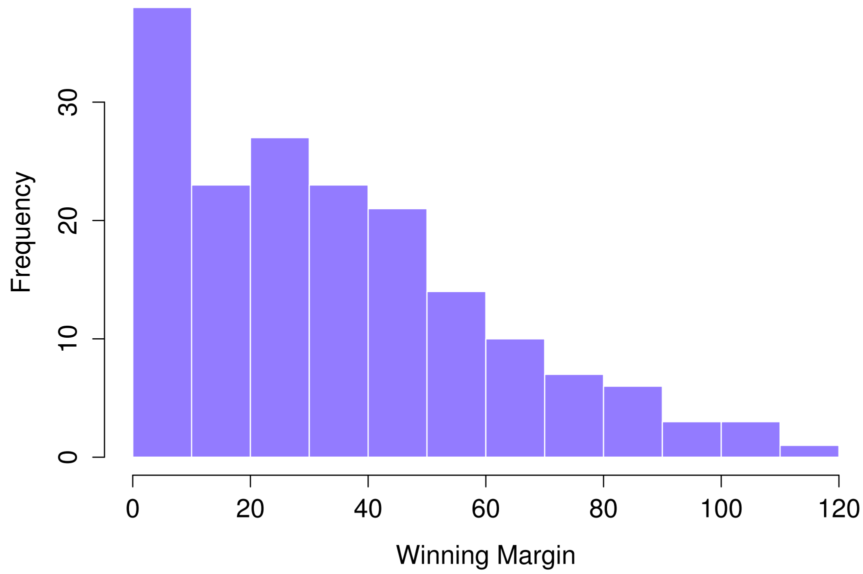


Figure1.2　2020 年の AFL 得点差データ (変数 afl.margins) のヒストグラム．ご想像の通り，より大差がつくゲームはより少ないのが見て取れます。

1.1.1　平均値

観測値のセットの**平均値**は，普通の，昔ながらの平均値です。全ての値を足し上げて，足した値の数で割ります。最初の 5 つの AFL の得点差は，56,31,56,8,32 ですが，これらの平均値を計算するには単に次のようにするだけです。

$$\frac{56+31+56+8+32}{5} = \frac{183}{5} = 36.60$$

もちろん，この平均の定義は誰にとっても新しいものではないでしょう。アベレージ (すなわち平均値) は，日常生活でもよく使われていますから，みなさんにとってもなじみ深い物でしょう。平均の概念についてはみなさん理解しているでしょうから，この計算を表記するために統計学者が使う数学的表記法について説明する機会とさせてもらって，その後で JASP でどのように計算するか紹介することにしましょう。

最初に導入する表記法は N です。これは平均するときの観測度数の数を表すのに使います (今回の場合は $N = 5$ です)。つぎに，観測値そのものについてのラベルをつけます。これには伝統的に X が用いられ，具体的にそのどれを指し示すのかについて，添字を使います。つまり，X_1 とすれば最

初の観測値，X_2 とすれば 2 番目の観測値，以下同様に X_N までいきます。あるいは，同じことをもう少し抽象的に表現するために，X_i で i 番目の観測値を指すことにします。表記法についてはっきりさせるために，以下の表では `afl.margins` 変数にある 5 つの観測について，数学的表記法と対応する実際の値の関係をリストアップしています。

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

オウケイ，では平均の式を書いてみましょう。伝統的に，平均を表すのに $\bar{X}$ を使います。平均の計算は以下の式で表現できます。

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{N-1} + X_N}{N}$$

この式はまったく正しいのですが，ちょっとばかり長ったらしいので，**総和記号**である $\sum$ を導入してこれを短縮しましょう[a]ここでは最初の 5 つの観測について足しあわせをしたいわけですから，長い書き方ですと $X_1 + X_2 + X_3 + X_4 + X_5$ となりますが，ここで総和の記号を使ってこれを次のように短縮します。

$$\sum_{i=1}^{5} X_i$$

文字通り，これは「1 から 5 までの全ての i について，X_i の値を足し合わせる」と読みます。しかしその意味は基本的に「最初の 5 つの観測値を足す」，です。どちらにせよ，これは平均を使うための記号として使われ，次のように書きます。

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

正直なところ，この数学的な表記法が平均の概念を明確にするのに役立つとは思えません。実際には，私が言葉で言ったのと同じことを書き出しているだけです。すなわち，全ての数字を足しあわせて，足した項の数で割る，です。しかし詳細に書き込んだ本当の理由はこれではありません。私のゴールは，誰もがこの本を読むときに使われるであろう記号について，はっきりと理解しておいてもらうことにあります。$\bar{X}$ は平均，$\sum$ は総和，X_i は ith 番目の観測値で N は観測の総数，ということをね。これらの記号は再利用されるので，みなさんがこれを使った式を「読む」ことができるように，さらに「多くのものを足しあわせて別のもので割る」と言えるように理解してもらうことが重要なのです。

[a] 総和に対して Σ を使うのは，勝手に決めたわけではありません。これはギリシア文字シグマの大文字で，アルファベットで言う S のアナロジーだからです。同様に，全ての積を示すための記号もあって，それは “products”(総積) と呼ばれるので文字としては Π を使います (ギリシアのパイの大文字で，これはアルファベットの P のアナロジーだからです)。

1.1.2 JASP での平均の計算

数学の話はここまで。計算してくれる魔法の箱はどうやって手に入れたらいいでしょうか？ 観測値の数が大きな数字になったら，コンピュータを使って計算させるのが何より簡単です。全てのデータを使って平均の計算をするために，JASP を使いましょう。最初のステップは ‘記述’ のボタンをクリックして，次に ‘記述統計’ をクリックしてください。それから変数 `afl.margins` をハイライト

させて，'右矢印' をクリックしてそれを '変数ボックス''に移します。するとすぐに画面の右側に表が現れます。そこには '記述' についての情報があります。図**??**を見てください。

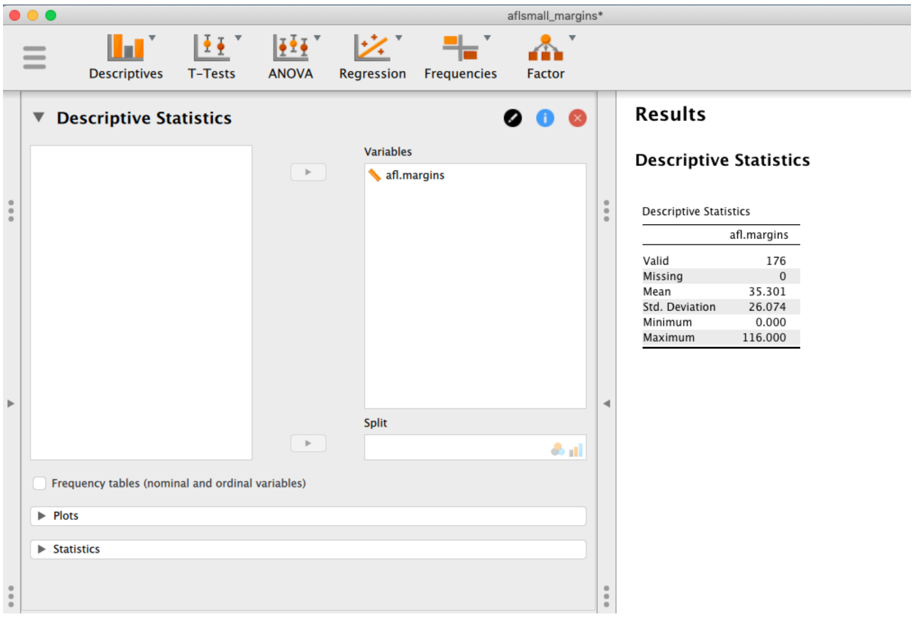


Figure1.3　AFL における 2010 年得点差データ (変数 `afl.margins`) のデフォルトで示される記述統計

図 **??**に見て取れるように，変数 `afl.margins` の平均値は 35.301 です。他の情報として，観測度数の総数 (N=176) や，欠損値の数 (ありません)，変数の中央値，最小値，最大値も含まれていますね。

1.1.3　中央値

中心化傾向の二つ目の測度としてよく使われるのは，**中央値**です。これの説明は平均よりも簡単です。変数セットの中央値というのは，ちょうど真ん中の値という意味です。AFL データの最初の 5 つの値，56,31,56,8,32 に興味があると思ってください。これらの数字の中央値を探すために，これを昇順に並べ替えます。

$$8, 31, \mathbf{32}, 56, 56$$

見てみると，これら 5 つの観測値の中央値は 32 ですね。並べ替えたリストの真ん中にあるからです (より分かりやすくするために，太字にしました)。簡単なことです。でも 5 つでなくて 6 つの観測値に興味があったらどうしましょう?シーズン 6 番目のゲームが得点差 14 点だったとすると，並べ替えリストは今や次のようになります。

$$8, 14, \mathbf{31}, \mathbf{32}, 56, 56$$

そして真ん中の数字は ふたつあって，31 と 32 になります。中央値は，この二つの数字の平均値として定義されるので，31.5 になります。前と同じで，数字がもっとたくさんあると人の手でやるのはとても難しくなります。実際には，もちろん，誰も真ん中の値を探すためにデータを並べ替えるなんてことはしません。コンピュータを使って，この面倒な作業をやらせるのです。JASP はお願いしたら中央値を出してくれます; 単に '統計' をクリックして，ドロップダウンメニューから '中心化傾向' メニューの '中央値' を選んでください。結果は自動的に中央値を含むものにアップデートされ，JASP は `afl.margins` 変数の中央値が 30.500 であるとレポートしてくれます。

1.1.4 平均値か中央値か?その違いは?

平均値と中央値の計算方法を知ることは，このお話の一部に過ぎません。あなたはそれぞれがデータの何についてものを言い，それらを使うときに何が仄めかされることになるのかを理解する必要があります。図**??**にそれを描いてみました。平均は，データセットの "重心" のようなもので，中央値はデータの "真ん中の値" です。これが意味することは，あなたがこれらのどちらかを使うときに，データの種類が何であって，それで何をやろうとしているのかに関わってきます。ざっくりいうと，

- データが名義尺度水準であれば，平均値も中央値も使うべきではありません。平均値も中央値も数字が割り当てられた値に意味がある，という考え方に依存しているからです。
- データが順序尺度水準であれば，平均値よりも中央値を使う方が良いでしょう。中央値はあなたのデータの順序情報 (すなわち，どの数字が大きいか) にだけ関わり，正確な数字には依存しないからです。これこそあなたのデータが順序尺度水準である状況でしょう。それに対して平均は，正確な量的値が観測対象に割り当てられているときに使われるので，順序尺度データには適していないのです。
- 間隔尺度あるいは比率尺度水準のデータであれば，どちらでも一般的に受け入れられます。どちらを選ぶかは，あなたが何をしたいかによります。平均値はデータの全ての情報を使います (あなたが大量のデータを持っているときには便利です)。が，極端な，外れ値には敏感です。

最後のパートを少し拡張しましょう。一つの結論として，平均値と中央値の間の体系的な違いは，ヒストグラムが非対称であるとき (歪んでいるとき；セクション **??**を参照) に現れます。これは図 **??**に描かれています。中央値は (右図)，ヒストグラムの "ボディ' ' 近くにありますが，平均値 (左図) は "尻尾"(極端な値があるところ) に引っ張られています。わかりやすい例を示すために，ボブ (年収$50,000)，ケイト (年収$60,000)，ジェーン (年収$65,000) が席についていると思ってください。テーブルの平均値は$58,333 で，中央値は$60,000 です。ここにビルが座ります。彼の年収は ($100,000,000) です。年収の平均値は$25,043,750 に跳ね上がりますが，, 中央値は$62,500 にあがるだけです。席についている人の全体的な年収に興味があるおなら，平均が正しい答えになるでしょう。しかし典型的な年収の人が知りたいのであれば，中央値がより良い選択肢になるのです。

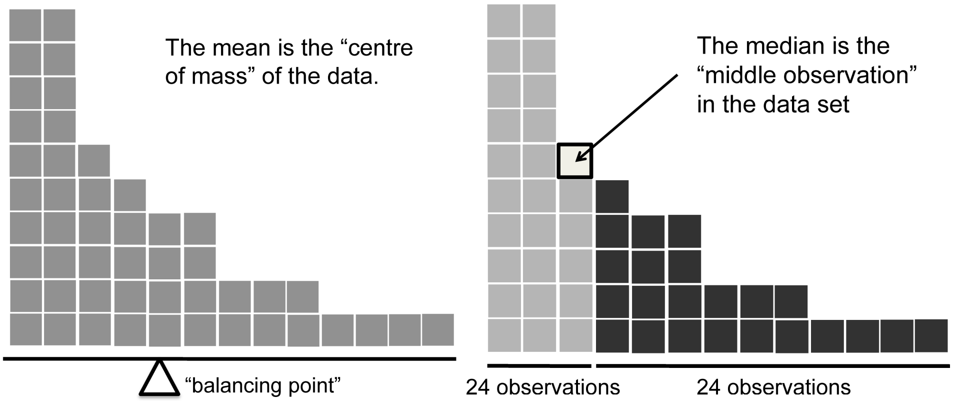


Figure1.4　平均値と中央値の違いをどう解釈するかについてのイラスト。平均値は基本的にデータセットの“重心”です。データのヒストグラムが固形物だと考えたら，そのバランスを取る点(シーソーみたいに)が平均値です。それに対して，中央値は真ん中の観測で，それより小さいデータが半分，それより大きいデータが半分あるということです。

. . . .

1.1.5　現実的な例

平均値と中央値の違いについて，何故注意を払うべきなのかの感覚を得るために，現実生活での例で考えてみましょう。私は科学的・統計的知識の足りないジャーナリストを馬鹿にする傾向があるのですが，信頼すべきところは信頼すべきだと思っています。これは 2010 年 9 月 24 日の ABC ニュース[*2] になった，ある素晴らしい論文です。

> コモンウェルス銀行の上級幹部がこの数週間、世界各地を訪問し、オーストラリアの住宅価格と所得に対する主要な価格の比率が、類似国と比較してどのように優れているかを示すプレゼンテーションを行いました。“住宅価格はこの 5,6 年，実質的に横ばい状態である” と銀行トレーディング部門のチーフエコノミスト Carig James は言っています。

これはおそらく、住宅ローンを抱えている人や、住宅ローンを希望している人、家賃を払っている人、オーストラリアの住宅市場でここ数年続いていることに全く気がついていない人にとっては、大きな驚きではないでしょうか。元の論文に戻ってみましょう。

> CBA(コモンウェルス銀行のこと) は，グラフ，数字，国際比較などで住宅の運命が決まると信じている人と戦ってきました。プレゼンテーションの中には，オーストラリアの家賃は収入に比べて割高であるという議論を，銀行が否定しているとされています。オーストラリアにおいて，世帯主の価格に対する住宅価格は大都市において 5.6，全国的には 4.3 であり，他の多くの先進国と同じぐらいであるとしています。また，サンフランシスコとニューヨークではこの比率は 7，オークランドでは 6.7，バンクーバーでは 9.3 にもなります。

もっとびっくりなニュースです! だけど，この論文は次のように見立てています。

> アナリストの多くは，これは銀行によってミスリーディングな図，比較がなされたからだと言い

[*2] www.abc.net.au/news/stories/2010/09/24/3021480.htm

> ます。CBA の資料 4 ページ目をみて，グラフや表の下に書いてある情報ソースをみたら，国際比較の追加的なソースがあることに気づくでしょう–人口動態学についての。コモンウェルス銀行が人口動態学の情報を使ってオーストラリアの住宅価格・収入比率の分析をしていたとすると，その実態は 5.6 とか 4.3 ではなく 9 近くになります。

うーむ，かなりの違いがありますね。一方では 9 といい，他方では 4-5 だ，と言っています。この違いを区分して，本当の値はこの間にあるんだとでもしたほうがよいでしょうか? 全く違います! 正しい答えと，間違った答えがあるような状態なのです。人口動態学は正しく，コモンウェルス銀行は間違っています。論文では次のように指摘しています。:

> コモンウェルス銀行の住宅価格対収入の図には明らかな問題があり，平均年収と住宅価格の中央値を比較しているのです (人口動態学の図は収入の中央値と価格の中央値の比較をしているのに)。中央値は真ん中にある点で，極端に高いあるいは低い値を効率よくカットしますが，平均値は年収や資産価値については高所得者が含まれるので高くなる傾向があります。別の言い方をすれば，コモンウェルス銀行の図は Ralph Norris の数百万ドルにも及ぶ給料を収入が話に入れ，かれの (間違いなく) 高価な住宅価格は図の中に入れないようにしているので，住宅価格はオーストラリアの中級ぐらいの年収と比較していることになります。

これ以上いうことはありません。人口動態学的に計算した比率の方が正しいのです。銀行がやったやり方は間違っています。なぜ数字に得意なはずの銀行がこのような基本的なミスをしたのかというと... 彼らが何を考えていたのかは分からないので，ここまでにしましょう。しかしこの論文が以下の事実についての注意を促しています。関係があるかどうかわかりませんが。

> オーストラリア最大の住宅業界牽引者であるコモンウェルス銀行は，住宅価格の上昇については最大級の興味を持っています。住宅ローンや多くの中小企業向けローンの担保として，オーストラリアの住宅の大部分を事実上所有しています。

むにゃむにゃ。

1.1.6 最頻値

サンプルの最頻値は，とても単純です。それは最も頻度が多い値，なのです。AFL の別の変数を使ってこれを説明してみましょう。決勝で最も多くプレーしている選手は誰でしょう？ `aflsmall_finalists` ファイルを開いて，`afl.finalists` 変数をみてみましょう。図 **??**がそれです。この変数には全 400 チームの，1987 年から 2010 年までの間に開催された 200 回の決勝戦情報が載っています。

我々がやる べきことは，全 400 試合を読み通して，決勝戦リストに出てくるチームの名前を数え上げ，**度数分布表**を作ることです。しかしこれは頭を使わない退屈な作業で，まさにコンピュータが得意とするような作業ですね。だから JASP を使いましょう。'記述' の下にある '記述統計' の，`afl.finalists` 変数を選び '変数' ボックスに移し，'度数分布表' と書かれた小さなチェックボックスをクリックします。すると図 **??**のようなものが得られるでしょう。

さて度数分布表を入手したわけですが，これをみると 24 年間ずっと，Geelong が他のどのチームよ

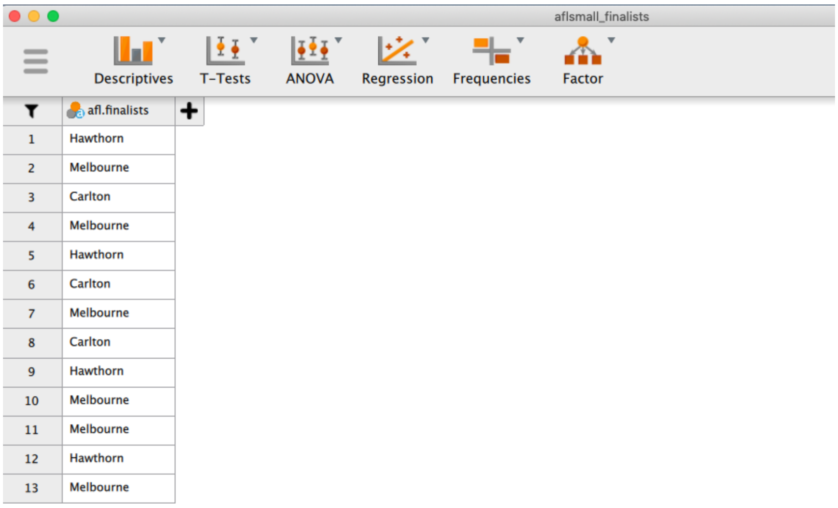


Figure1.5　`aflsmall_finalists.csv` ファイルに修められた変数の JASP スクリーンショット

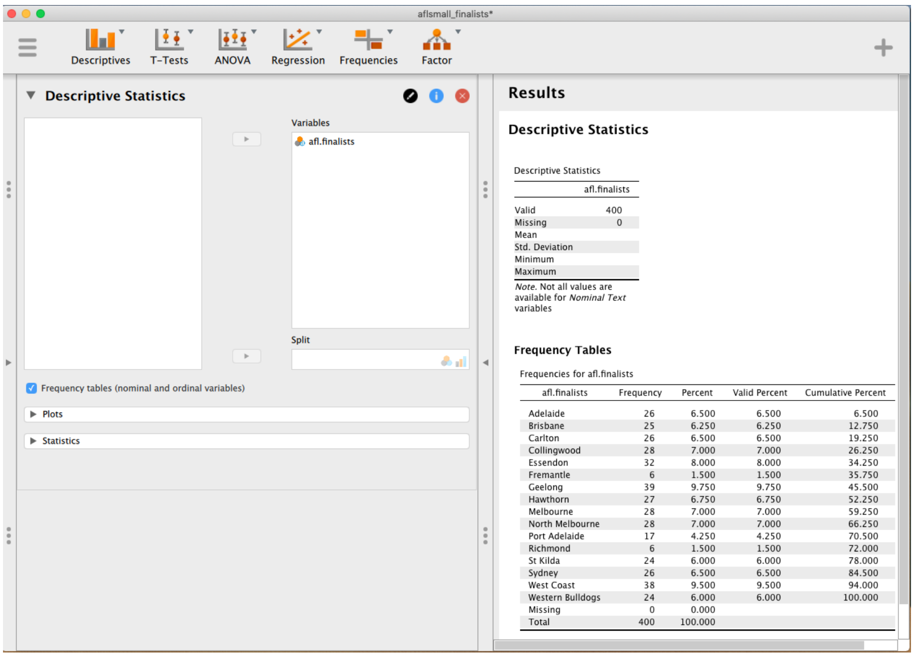


Figure1.6　`afl.finalists` 変数の度数分布表を示した JASP スクリーンショット

りも多く決勝に進んでいることがわかります。ですから `afl.finalists` データの最頻値は`"Geelong"`だということになります。Geelong(39 回決勝進出) が 1987 年から 2010 年の間で他のどのチームよりも多く決勝に進んでいるのです。また，'記述統計' の表では平均値，中央値，最大値，最小値が計算されていないのも注目です。なぜなら `afl.finalists` 変数は名義的な文字変数であって，これらの値を計算する意味がないからです。

最後に最頻値に関するポイントをもう一つ。名義尺度のデータを持っていたら最頻値を計算するのが最もよくあるケースです。というのも，平均や最頻値はこの種の変数には向いていないからですが，順序，間隔，比率尺度水準の変数の最頻値を知りたいという時もあります。例えば，`afl.margins` 変数にもどってみましょう。この変数は明らかに比率尺度水準 (もしピンとこないのなら，もう一度セクション **??**を読んでみてください) であり，あなたが知りたいのはこの中心に関する測度であれば平均値や中央値を求めるところです。しかしこんなことを考えてみてください：あなたの友達が賭けようぜと言ってきて，ランダムにフットボールのゲームを選ぶとします。誰がプレイするのかを知らずに，*正確な* 得失点差を推測しないといけないのです。正しく当てられたら 50 ドルもらえます。でなければ 1 ドル失います。ほとんど正解に近かった，という残念賞はないものとします。正確に点差を推測しなければならないのです。この賭けをする時，平均や中央値は全くあなたの役に立ちません。最頻値にかけるべきです。`afl.margins` 変数の最頻値を JASP で計算するには，データセットに戻って '記述'-' 記述統計' 画面から，'統計量' と書いてあるセクションを拡大してください。'最頻値' のチェックボックスをクリックして，'記述統計量' テーブルにある最頻値をみます。図**??**にあるやつです。そうすると，2010 年のデータでは 3 点差に賭けるべきだということがわかります。

1.2 変動性の指標

ここまで話してきた統計の話は，*中心化傾向*に関するものでした。つまり，そこでの話はデータの "真ん中" とか "代表的な" 値についてでした。しかし，中心化傾向は計算したい要約統計量の唯一の種類，というわけではありません。計算したい第二のものとして，データの**変動性**があります。つまり，どれぐらいデータが "散らばっているか"? とか，どれぐらい平均や中央値から観測値が "遠くにある'' 傾向があるか?というものです。ここでは，データが間隔あるいは比率尺度水準で得られていると考えますから，`afl.margins` データを例に使い続けましょう。このデータを使うことで，散らばりの指標としていくつかのものを示すことにし，その長所と短所も見ていくことにしましょう。

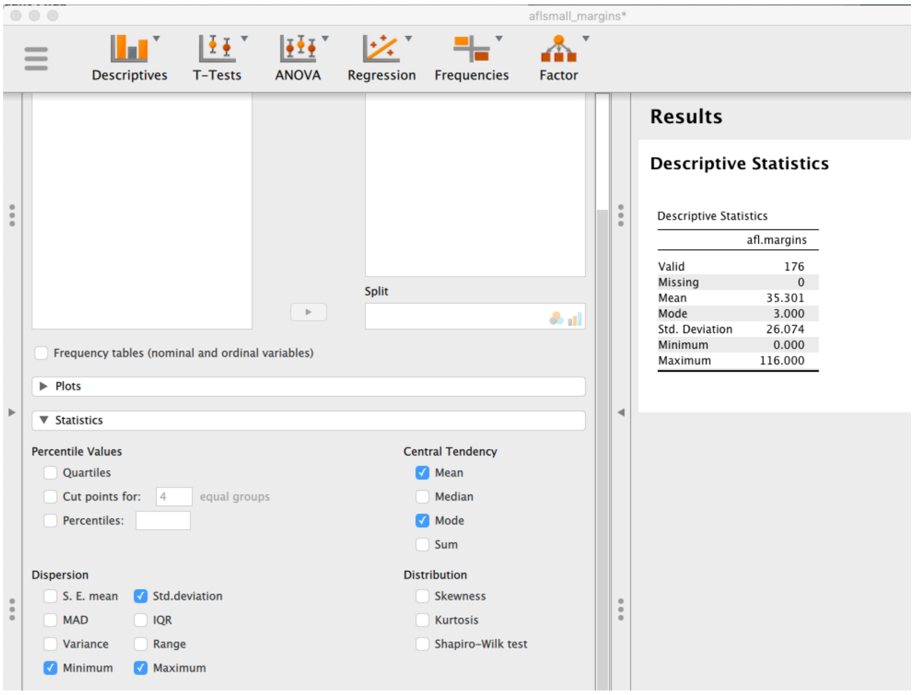


Figure1.7　`afl.margins` 変数の中央値を示す JASP 画面

. . . .

1.2.1　範囲

変数の**範囲**はとてもシンプルなものです。最大値から最小値を引いたもののことを指します。AFL 得点差データの最大値は 116 で最小値は 0 でした。“変動” を表す量として範囲は最も単純なものですが，最も悪いものでもあります。要約統計量を頑健なものにするために，平均について議論していたことを思い出してください。もしデータセットの中に一つ二つ変な値があると，我々の統計量はそうしたデータに角に影響されないようにしたいところです。

例えば，変数が極端な外れ値を持っていたとします。

$$-100, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

範囲が頑健な値ではないことは明らかですよね。変数の範囲は 110 になりますが，外れ値を除くとたったの 8 になります。

1.2.2　四分位範囲

四分位範囲 (interquartile range,IQR) は範囲に似ていますが，最大値と最小値の差を使うのではなく，25 パーセンタイルと 75 パーセンタイルの差を使います。**パーセンタイル**をまだ知らないかもし

れませんが，データの 10 パーセンタイルというのはある点 x よりも小さいのがデータの 10% になるような点 x のこと，という意味です。実は，既にこの考え方は出てきているのです。データの中央値とは，50 パーセンタイルのことですから! JASP では，簡単に 25,50,75 パーセンタイルを見つけることができます。‘記述’ の ‘記述統計’ から ‘統計量’ の画面にある ‘四分位’ チェックボックスをクリックするだけです。

Descriptive Statistics

Descriptive Statistics

	afl.margins
Valid	176
Missing	0
Mean	35.301
Mode	3.000
Std. Deviation	26.074
Minimum	0.000
Maximum	116.000
25th percentile	12.250
50th percentile	30.500
75th percentile	51.500

Figure1.8 `afl.margins` 変数の四分位を示す JASP のスクリーンショット

. . . .

驚くには値しませんが，図**??**において 50 パーセンタイルは中央値と同じになっています。そして，$50.50 - 12.75 = 37.75$ ですから，2010 年の AFL 得点差データの四分位範囲は 37.75 ということになります。範囲の解釈は明らかですが，IQR の解釈の仕方はそこまで明らかだというわけではないですね。これは次のように考えるのが最も単純な方法でしょう。すなわち，四分位範囲はデータの “中半分の” 範囲だというものです。つまり，データの一つの四分位が 25 パーセンタイル点で，もう一つの点が 75 パーセンタイル点ですから，この二つの間にデータの “中半分” が位置していることになります。IQR はこの中半分をカバーする範囲なのです。

1.2.3 平均絶対偏差

二つの尺度，範囲と四分位範囲を’ みてきましたが，どちらもデータのパーセンタイルをみて，データの散らばりを測ろうとするアイデアに基づいているものでした。しかし，これだけがこの問題唯一の解決策ではありません。別のアプローチとして，意味のある参照点 (ふつう平均値や中央値ですが) を選び，その参照点からの “典型的な” 偏差を報告する，というのがあります。“典型的な”

偏差，というのは何を意味しているでしょう？ 普通これは偏差の平均値や中央値を指します。実際，ここからは二つの尺度が導かれます。“平均絶対偏差”(平均値からの) と，“中央値絶対偏差”(中央値からの)，です。私がこれまでみてきたところ，中央値に基づく尺度が統計的に使われているようで，そちらの方が優れているようです。しかし正直に言って，心理学でこれらが使われてきたのをあまりみたことがありません。平均に基づく尺度の方が，心理学ではよく出てきます。このセッションでは前者について最初説明しますが，その後で 2 番目についても触れていきます。

前のパラグラフではちょっと抽象的だったかもしれませんが，平均からの**平均絶対偏差**についてもう少しゆっくりみていきましょう。この尺度が便利なことの一つに，この名前が実際にどうやって計算するのかを表している，というのがあります。AFL の得点差データについて，もう一度最初の 5 ゲームをみてみると，得点差は 56，31，56，8，32 でしたね。ここでの計算はある参照点 (今回は平均) からの偏差を見るものですから，最初にするべきことは平均つまり $\bar{X}$ を計算することです。最初の 5 ケースでは，平均は $\bar{X} = 36.6$ になりました。次のステップは各観測値，X_i を偏差のスコアに変換することです。これは観測値 X_i と平均 $\bar{X}$ の差を計算することでできます。つまり，偏差スコアの定義は $X_i - \bar{X}$ となるのです。今回のサンプルにおける最初の観測値は，$56 - 36.6 = 19.4$ になります。オーケイ，十分シンプルですね。このプロセス，次のステップはこれらの偏差を絶対偏差にすることです。これは負の値を正の値にすることでできます。数学的には -3 の絶対値を $|-3|$ と書き，$|-3| = 3$ とします。この絶対値を使うのは，平均よりも高かったのか低かったのかを気にしないということであり，興味は平均にどれぐらい *近かったのか*というだけだということです。このプロセスをできるだけ明白にするために，下の表では，5 つの観測値すべてについてのこれらの計算を示しています。

用語:	どのゲームで	値	平均偏差	絶対偏差
表記:	i	X_i	$X_i - \bar{X}$	$\lvert X_i - \bar{X}\rvert$
	1	56	19.4	19.4
	2	31	-5.6	5.6
	3	56	19.4	19.4
	4	8	-28.6	28.6
	5	32	-4.6	4.6

さてデータセットの各観測値について絶対偏差を計算できたので，これらのスコアの平均を計算しましょう。次のようになります。

$$\frac{19.4 + 5.6 + 19.4 + 28.6 + 4.6}{5} = 15.52$$

はいおしまい。これら 5 つのスコアについて，平均絶対偏差は 15.52 でした。

ところで，この簡単な例はこれでおしまいですが，少し話が残っています。まず，数学的な定式化をしておくべきです。しかしこれをしようとすると，平均絶対偏差についての数学的表記が必要です。腹立たしいことに，“平均絶対偏差” と “中央値絶対偏差” はどちらも同じ頭文字 (MAD) なので，曖昧になってしまいますから，平均絶対偏差に何か別の表現を考えないといけないでしょう。やれやれ。*average* absolute deviation を短くして，AAD とすることにしましょう。これでもいくらか曖昧な表記ですが，計算は次のように書くことができます。

$$\text{AAD}(X) = \frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|$$

1.2.4 分散

平均絶対偏差は使いでがありますが，変動の尺度として最適というわけではありません。純粋に数学的な観点からは，絶対偏差よりも二乗した偏差の方が好ましい理由があります。これを使うと**分散**とよばれる尺度を手に入れることになります。それは本当にステキな統計的特徴を持っているのですが，それは横に置いておくとして[*3]，今から取り上げるとても大きな心理学的欠陥も持っていることを説明します。データセット X の分散は $\text{Var}(X)$ と表記されますが，もっと一般的には s^2 と書きます (その理由はすぐにわかります)。

観測されたデータセットの分散を計算する式は次の通りです。

$$\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$$

ご覧の通り，基本的には平均絶対偏差で使ったものと同じ形をしていますが，違うのは “絶対偏差” のかわりに “偏差平方” を使っているところです。このため，分散は “平均偏差平方” とも言われます。

さて，基本的な概念を手に入れましたので，具体例でみてみましょう。もう一度，AFL ゲームの最初の 5 つのデータを使います。前回同じアプローチをした時に習って，次のような表にしてみました。

[*3] えーっと，ちょっとだけ何が最高にクールなのか，“クール” の定義をしてから説明してみましょう。分散は *加算的* なのです。その意味はこんな感じです。私が二つの変数 X と Y を持っていて，それらの分散がそれぞれ $\text{Var}(X)$ と $\text{Var}(Y)$ だとしましょう。ここで新しい変数 Z を，二つの和，$Z = X + Y$ で定義したとします。そうすると，Z の分散は $\text{Var}(X) + \text{Var}(Y)$ になるのです。これが とても便利な特徴なのですが，このセクションで私が説明しようとする他の尺度にはないものなのです。

用語:	どのゲームで	値	平均偏差	偏差平方
表記:	i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	1	56	19.4	376.36
	2	31	-5.6	31.36
	3	56	19.4	376.36
	4	8	-28.6	817.96
	5	32	-4.6	21.16

最後の列には全ての偏差平方が入っていますので，これの平均を取れば良いのです。手計算する，つまり電卓を使うと，この分散の値が 324.64 であることがわかります。興奮してきたでしょう?このとき，多分あなたの考えに火がついた問題 (すなわち，324.64 の分散って本当に平均なのか?) は横に置いて，JASP でこれをどう計算するかをみてみましょう。というのも，これで奇妙なことが明らかになるからです。

まず最初の 5 行**だけ**を含んだ新しいデータを読み込みます。ファイル `aflsmall_margins_first5.csv` を読み込んでください。次に '統計' メニューの '記述'–'記述統計' をクリックし，'分散' チェックボックスをクリックします ('ばらつき' グループの中にあるのがわかると思います)。手計算した値 (`324.64`) と同じ数字になりましたか？ いや，ちょっと待って，あなたは全く *違う*答えを手にしたではありませんか (`405.800`)!おかしいなあ。JASP は壊れてるの?タイポですか?何が起こってる?

起こった通りのことで，答えは no です。タイポではなく，JASP が間違っているわけでもありません。現に，JASP がここで何をしているのかを説明するのはとても簡単なのですが，JASP が *なぜ* それをしたのか，というのはちょっと説明に苦労します。ですから "何が起こったのか" から始めましょう。JASP は上で示したのとは少し違う数式を評価したのです。偏差平方の平均を計算したのではありません。平均はデータ点の数 N で割りますが，JASP は $N - 1$ で割ったのです。

言い換えると，JASP は次の式を使って計算したのです。

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

これが *何をやったか*です。本当に知りたいのは，*なぜ* JASP が N ではなく $N - 1$ で割ったのか，ですよね。結局のところ，分散は偏差平方の *平均*なのですよね?だったら N で割るべきじゃないか，サンプルの実際の観測数でね。全くその通りです。しかし，第 **??**章で論じるように，"サンプルを記述する" ことと "サンプルのもとになった母集団を推測すること" とのあいだにはちょっとした違いがあるのです。ここまでは，この差の区別をしてきませんでした。あなたが表現したいのがサンプルなのか，母集団の推測するものなのかどうかにかかわらず，平均は同じように計算できたのです。しかし分散や標準偏差，そのほかの尺度ではそうならないのです。私が最初に説明したこと (つまり，

N で割ることによる実際の平均) は，標本の分散を計算することを想定したものでした。しかしほとんどの場合，標本 そのものに興味をもってるわけではないでしょう。むしろ，その標本は世界について何かを伝えるために存在しているはずです。そうであれば，あなたが実際に計算したいのは “標本統計量” ではなくて，“母集団の母数” を推定するためのものになるはずです。しかしこの話は，少し先走りすぎています。今は，JASP がすることをただ信じて，第 **??**章で推定について論じるときまでこの問題をおいておくことにしましょう。

最後にもう一つ。このセクションはちょっとした推理小説のようになっていました。先ほど分散の式を示し，JASP では “$N-1$” でやっていること，そしてなぜそうするのかのヒントを書きましたが，最も大事なことは触れていなかったのです。みなさんは分散をどういうものだと *理解*していますか?記述統計は記述することだけを目的としていますが，今のところ分散は意味不明な数字でしかありません。残念なことに，分散の解釈について人間味のない説明しかできない理由は，それがそもそも人間味のないものだからです。これが分散について最も深刻な問題点です。分散は本当は変動を表現する基本的な量であるというある種の美しい数学的特性はあるのですが，現実的に他者との会話に使いたいと思うときには全く役に立たないのです。分散は元の変数に関しては全く意味のない数字になります! 全ての数字は二乗されてしまうので，それは何も意味しないことになるのです。これは大問題だ。例えば，以前示した表について言うと，ゲーム 1 における点差は “376.36 ポイントの二乗分，平均より高い” と言うことになります。これは *まったく馬鹿馬鹿しい表現*ではないですか。計算した分散の 324.64 の時も同じことがいえます。多くのフットボールゲームを見てきましたが，誰も “ポイントの二乗分” なんて言ってるのを聞いたことがありません。これは測定の実際の単位では *なく*，分散は意味のない単位を持っているので，人間にとって全く意味のないことになるのです。

1.2.5　標準偏差

オーケイ，分散を使う理由は分かってもらえたとしましょう。説明はしてませんが，分散は数学的に良い特性持っていますからね。でもあなたが人間で，ロボットでないなら，データと同じ単位を持っている (つまり二乗した値ではないもの) 尺度を使う方がいいと思うでしょう。じゃあどうしましょう?答えは簡単です!分散の平方根を取れば良くって，これは**標準偏差**として知られています。“偏差平方平均の根”，つまり RMSD とも呼ばれます。これで問題がスッキリ解決しました。だれも “分散は 324.68 ポイントの二乗” ということの意味を理解することはできませんが，“標準偏差 18.01 ポイント” は簡単に理解できます。元の単位で表現されているんですから。

標準偏差は分散の正の平方根に等しいので，次の式を見ても驚かないと思います。

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

JASP では，'分散' のチェックボックスと同じセクションに '標準偏差' のチェックボックスもあります。図??をみると，JASP は `afl.margins` の標準偏差を `26.074` と答えてくれています。標準偏差はとてもよく使われるので，チェックするのがデフォルトになっていますが，あなた自身で選んでみてください!!

しかし，分散についての議論でお気づきかもしれませんが，JASP は実際にはこれとちょっと違ったやり方で計算します。分散を見るだけなら，JASP は N ではなく $N-1$ で割る方で計算するのです。

第 ??章で再びこのトピックに触れると意味がわかると思いますが，この新しい量を $\hat{\sigma}$("シグマ・ハット" と読みます) とし，次のように定式化します。

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

標準偏差を解釈するのも少し複雑です。標準偏差は分散から導出されています。そして分散は人にとってあまり意味のない量になっていますから，標準偏差は単純な解釈では済みません。結果的に，私たちのほとんどはちょっとした経験則を用いています。一般的に，平均から標準偏差 1 つぶん離れたところにデータの 68% が含まれ，データの 95% は平均から標準偏差 2 つ分離れたところに 99.7% が，平均から標準偏差 3 つ分離れたところに含まれる，ということが期待できます。このルールはほとんどの場合うまく当てはまりますが，多少の例外はあります。これがちゃんと計算できるのはヒストグラムが対称的で "ベル型" になっているという *仮定*に基づいています[*4]。図 ??にある AFL の得点差ヒストグラムを見ると，この経験則は私たちのデータに合っているとは思えません! しかし大まかに合っているのです。AFL データの 65.3% が実際に平均から 1 標準偏差の範囲にあります。This is shown visually in Figure ??. このことは，図 ??に視覚的に示されています。

1.2.6 どの尺度を使いましょうか?

いくつかの範囲についての尺度を紹介してきました。範囲，IQR，平均絶対偏差，分散，標準偏差

[*4] 厳密にいうと，この仮定はデータが *正規分布*にしたがっているということで，この重要な概念については第?? 章で議論することになります。またこのことは本書で何度も何度も出てきます。

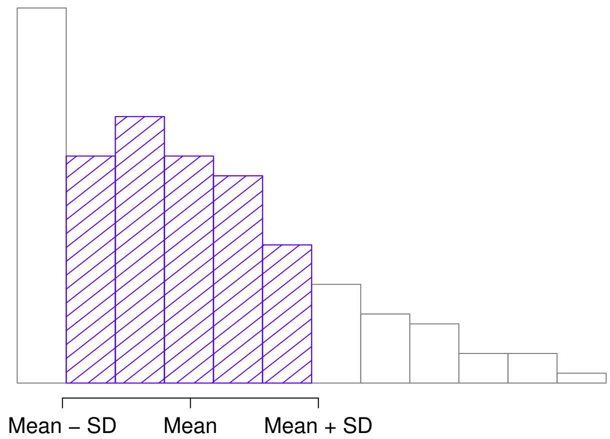


Figure1.9　AFL 得点差データについての標準偏差を描いたもの。色がついているヒストグラムの箇所は平均から 1 標準偏差のなかに入ったデータの数を表しています。今回は 65.3% のデータセットがこの範囲内に入り，次のメイントピックスである “約 68%’’ のルールに近い結果になっています。

....

です。そしてその長所と短所についてもみてきました。簡単にまとめておきましょう。

- *範囲* データのちらばり全体を見ます。外れ値に弱く，データの極端な部分を見たいという理由がない場合はあまり使われることはありません。
- *四分位範囲* データの “真ん中あたり” がある場所を教えてくれます。多少，外れ値に強くて中央値を含んでいます。これはよく使われます。
- *平均絶対偏差*平均から観測度数が “平均的に” どれぐらい離れているかを教えてくれます。解釈しやすいのですが，いくつかの小さな問題点があって (ここでは触れていませんが)，そのせいで統計家は標準偏差ほど魅力を感じていません。時々使われますが，それほど頻度は多くありません。
- *分散* 平均偏差の二乗の平均です。数学的にはエレガントで，平均周りの散らばりを描写するにはたぶん “正しい” 方法なのですが，データと同じ単位を使っていないので意味不明な数字になります。数学的なツール以外の用途はほとんどありませんが，非常に多くの統計技法の中に “埋もれて” います。
- *標準偏差*分散の平方根です。これは数学的にも非常にエレガントで，データと同じ単位で表現されていますから，解釈も簡単です。平均が中心化傾向の尺度として使われる時は，これが基本です。散らばりの尺度の中で最もポピュラーなものになります。

まとめると，IQR と標準偏差が簡単で，データのばらつきを報告するのに最もよく使われる二大尺度，ということになります。しかし他のものが使われることもあります。この本に載せたのは，わずかではありますがみなさんがどこかで出会うかもしれないからです。

1.3 歪度と尖度

みなさんが心理学の文献で見かけるかもしれない記述統計量が，あと二つあります。歪度と尖度です。実践上はどちらもこれまで話してきた中心化傾向や変動性の尺度ほど，使われるものでもありません。歪度はちょっと大事なので見かけることはあるかもしれませんが，私は科学的レポートの中で尖度を目にしたことはありません。

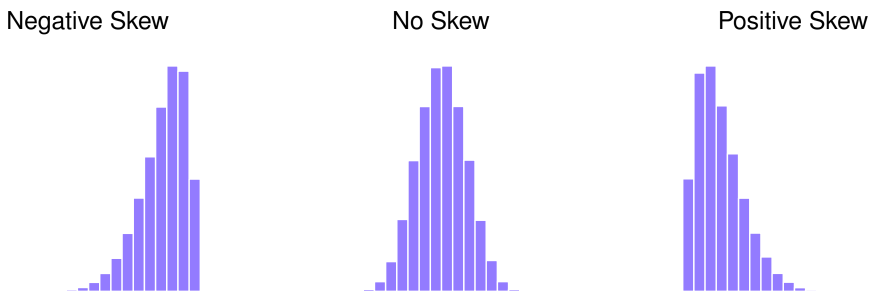


Figure1.10　歪度のイメージ。左側は歪度が負 (歪度 $= -.93$)，真ん中は歪みなし (実際ほとんどありません。歪度 $= -.006$)，そして右が正の歪度 (歪度 $= -.006$) をもつデータです。

歪度の方が面白いので，こちらから話を始めましょう。歪度は基本的に非対称性の尺度で，図を書いてみれば理解は簡単です。図 **??**にあるように，データに極端に小さな値 (下の裾が上の裾よりも“長い”) を持っていて，極端に大きな値はそれほど持っていない (左図) 場合，このデータは *負の歪度*をもつといいます。一方，極端に大きな値が小さい値より大きくも多くあるようであれば (右図)，このデータは *正の歪度*をもつといいます。これが歪度の背後にある考え方です。平均よりも大きな値が相対的に多くあれば，分布は正，すなわち右に歪んでおり，裾も右に寄っています。負，すなわち左への歪みはその逆です。対称的な分布をしていれば，歪み度は 0 です。正に歪んだ分布の歪度は正の値であり，負の値は負の歪み分布だと言えます。

データセットの歪みについての定式化は次のとおりです。

$$\mathrm{skewness}(X) = \frac{1}{N\hat{\sigma}^3}\sum_{i=1}^{N}(X_i - \bar{X})^3$$

ここで N は観測度数の数であり，$\bar{X}$ は標本平均，$\hat{\sigma}$ は標準偏差 (ただし “$N-1$ で割ったバージョン”) です。

ありがたいことに，JASP で歪度の計算をすることができます。‘記述’ - ‘記述統計量’ の下にある ‘統計量’ チェックボックスのオプションがそれです。変数 `afl.margins` について，その歪度を計算すると `0.780` です。この歪度の推定値を歪度の標準誤差で割れば，このデータがどれほど歪んでいるかの指標を得ることができます。経験的に行って，小さいサンプルでは ($N < 50$)，この値が 2 以下であればそれほど歪みは大きくなく，2 以上であればデータが統計的な分析をするに許される限界を超えて歪んでいる，と考えるのが目安です。これは経験則に過ぎず，この解釈にはっきりした共通見解があるわけではないことに注意してください。ということで，この分析をすると AFL の得点差データはちょっと歪んでいることになります (`0.780 / 0.183 = 4.262` で，これは明らかに 2 より大きいです)。

時々つかわれる最後の尺度は，実際に使われることは非常に稀なのですが，データセットの**尖度**です。簡単にいえば，尖度は “尖っているかどうか” の尺度で，図 **??**にその状況を示しています。慣例によって，“正規分布”(黒い線) は尖度ゼロであり，データセットの尖り具合はこのカーブに比べて相対的に評価されます。

この図にあるように，左のデータはそれほど尖っておらず，尖度は負でこのデータは *緩く尖った platykuritic* データだと言われます右図はとても尖っており，尖度は正でこのデータは *尖度の大きい leptokurtic* データだと言われます。一方，真ん中のデータはちょうどいいぐらいの尖度で，これは *中程度の尖度 mesokurtic* と呼ばれ，尖度はゼロです。下の表にこれをまとめました。

一般的な言い方	専門的な言い方	尖度の値
“かなりフラット”	platykurtic	負
“ちょうどいいぐらい”	mesokurtic	ゼロ
“とても尖っている”’	leptokurtic	正

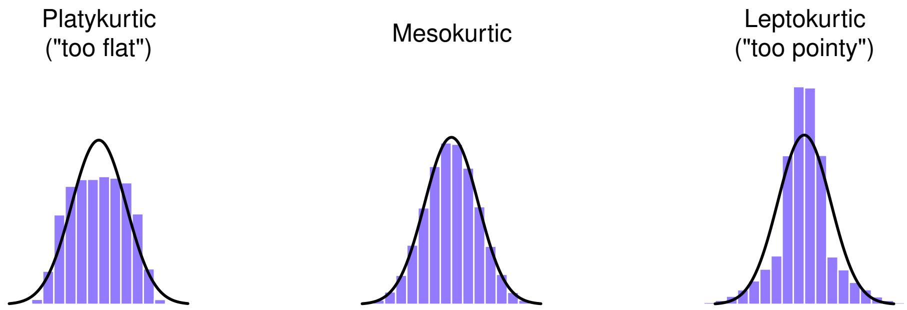


Figure1.11　尖度の図。左側は “緩く尖った” データセット (尖度 $= -.95$) であり，これが意味するのはこのデータセットは “かなりフラット’’ だということです。真ん中の図は “中程度の尖り” をもったデータセット (尖度はほとんど 0) であり，これが意味するのはこのデータの尖度がちょうどいい感じであるということです。最後に，右側の図ですが，“尖度の大きい” データセット (尖度 $= 2.12$) であり，このデータセットは “とても尖っています”。尖度は正規分布 (黒い線) と比べて評価されていることに注意してください。

. . . .

尖度の式は既に見た分散や歪度の式とかなり似ています。分散が偏差の二乗，歪度が偏差の三乗であったのに対し，尖度は四乗になっています。[a]

$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^4 - 3.$$

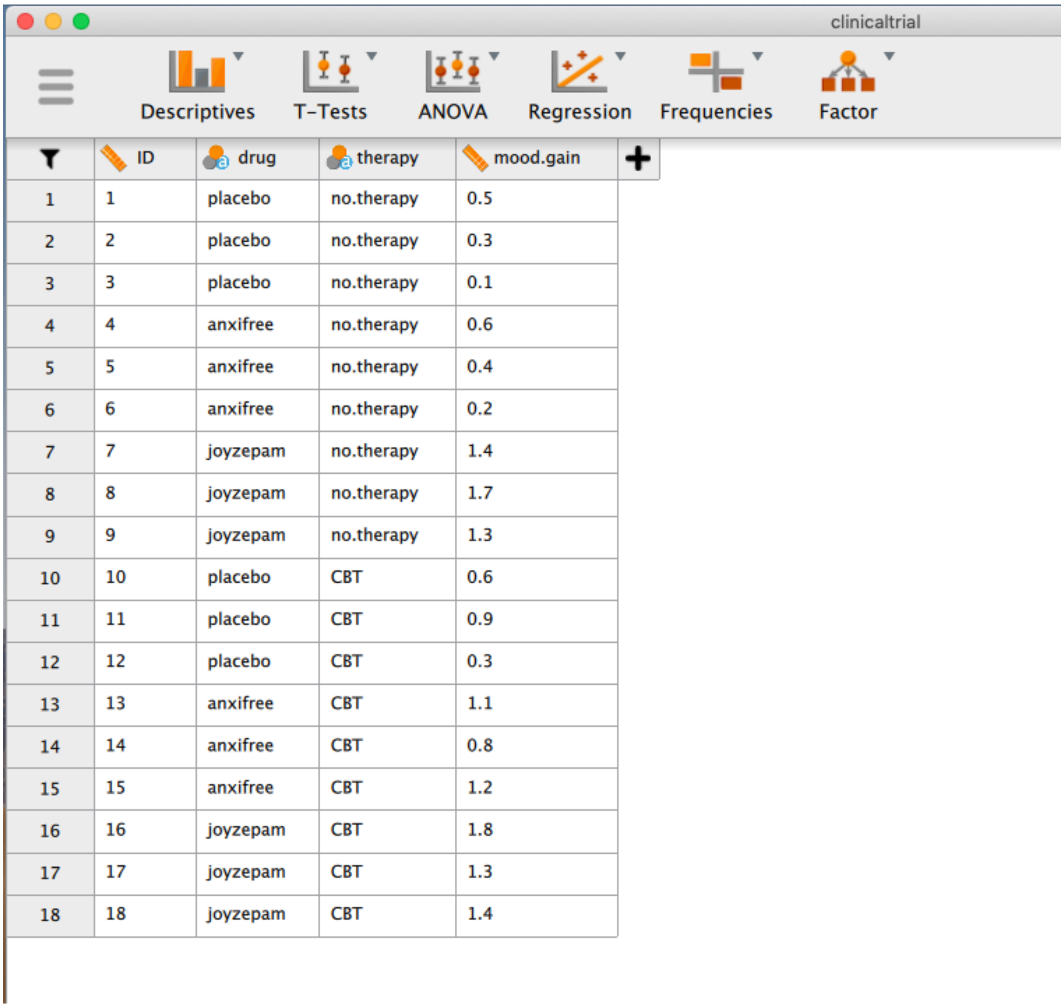
[a] この “-3” については正規分布の尖度がゼロになるように統計家が付け加えたものです。“-3” を式の最後に引っ付けておくのはちょっと馬鹿みたいですが，こうすることの数学的な理由があるのです。

大事なのは，JASP で尖度を計算するには歪度の下のチェックボックスをクリックするだけだということで，そうすると尖度の値 `0.101` がその標準誤差 `0.364` と共に表示されます。歪度をその標準誤差で割ったのと同じように計算すると，この値は 2 より小さい ($0.101/0.364 = 0.277$) ことがわかります。これは AFL の得点差データの尖度がちょうどいいぐらいだったことを意味しています。

1.4

グループごとの記述統計

よくあることのひとつとして，記述統計量をあるグループ変数ごとに分割してみたいと思うことがあります。JASP ではすごく簡単にできます。例えば，ある `clin.trial` データについて，`therapy` のタイプごとに記述統計量を見たいなと思ったとしましょう。これは今まで見せていない，新しいデータセットです。このデータセットは `clinicaltrial.csv` ファイルにあって，第 **??**章でよく使うようになります (このデータの詳細についてはその時に説明します)。読み込んで，見てみましょう。



	ID	drug	therapy	mood.gain
1	1	placebo	no.therapy	0.5
2	2	placebo	no.therapy	0.3
3	3	placebo	no.therapy	0.1
4	4	anxifree	no.therapy	0.6
5	5	anxifree	no.therapy	0.4
6	6	anxifree	no.therapy	0.2
7	7	joyzepam	no.therapy	1.4
8	8	joyzepam	no.therapy	1.7
9	9	joyzepam	no.therapy	1.3
10	10	placebo	CBT	0.6
11	11	placebo	CBT	0.9
12	12	placebo	CBT	0.3
13	13	anxifree	CBT	1.1
14	14	anxifree	CBT	0.8
15	15	anxifree	CBT	1.2
16	16	joyzepam	CBT	1.8
17	17	joyzepam	CBT	1.3
18	18	joyzepam	CBT	1.4

Figure1.12　`clinicaltrial.csv` ファイルにある変数を写した JASP スクリーンショット

. . . .

三つのドラッグがあるのがわかりますね。プラセボと，“anxifree” と “joyzepam” と呼ばれるものです。そしてそれぞれに 6 人割り当てられてます。そして 9 人が認知行動療法 (CBT) を受けていて，

9 人が心理療法は何も受けていない状態です。そして mood.gain 変数の '記述' をみてみると，ほとんどの人が気分の向上 (平均 = 0.88) を示していますが，この尺度が何なのかわからないままでは，それ以上のことは言えません。でも，それはそれでわるくないのです。全体的には何か勉強になった気になります。

さて，さらに他の記述統計量を見て行きましょう。こんどはセラピーのタイプごとに分けて。JASP で '統計量' オプションから標準偏差，歪度，尖度にチェックを入れます。同時に，therapy 変数を '分割' ボックスに入れます。すると図??のような結果が得られます。

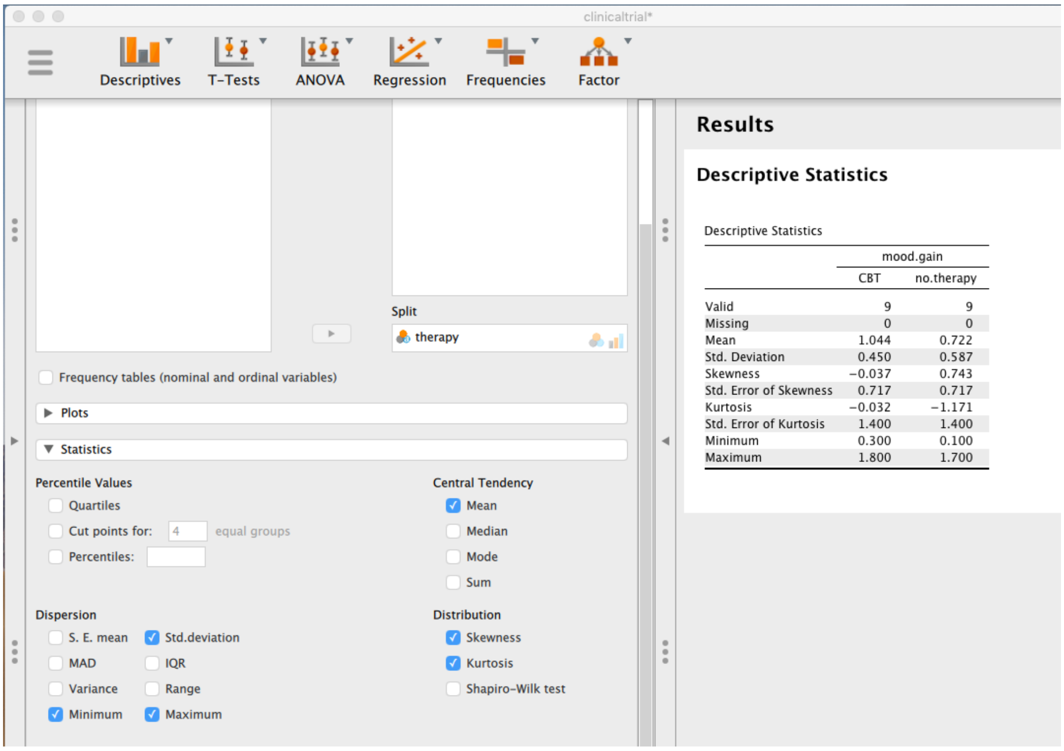


Figure1.13　セラピータイプごとに分割した記述統計量を示した JASP のスクリーンショット

1.5

標準得点

私の友人が " 不機嫌さ" を測定するための新しい質問紙を作ろうとしているとしましょう。この調査票は 50 の質問からなり，不機嫌かどうかについて答えるものとします。大きなサンプルをとって

(仮に百万人ぐらいとったとしましょう!)，このデータが正規分布しており，50 問中 17 点が平均不機嫌スコアで，標準偏差が 5 だとしましょう。これに比べて，私の得点は 50 問中 35 点だったとします。私はどれぐらい不機嫌なんでしょうね?これについて考える一つの方法は，私は 35/50 が不機嫌なのだから，70% ぐらい不機嫌だと考えることです。しかしちょっと考えてみれば，おかしい気もしますよね。もし私の友人が，その質問紙を少し違った捉え方で答えていたとしたら，その問いが本当に問うていることに比べて，全体的な分布が簡単に上がったり下がったりしてしまいます。ですから，私が 70% 不機嫌だというのは，*調査票の質問セットに応じて変わることになります。*とても良い質問項目であったとしても，これではあまり意味のある表現にはなりません。

これについての良いやり方の一つは，私の不機嫌の程度を周りの人と比べることです。驚くべきことに，私の友人は 1,000,000 人のサンプルを持っていて，その中でたった 159 人だけが私と同じ程度の不機嫌さ (本当ははそんなことありませんよ) であれば，私はトップ 0.016% の不機嫌度ということになります。このほうが，ロウデータを解釈しようとする時にはより意味があるのではないでしょうか。この考え方は，私の不機嫌さの程度を人の全体的な不機嫌分布にあわせて記述しようとするものであり，標準化がしようとしているのはまさにこれなのです。これを正しくやる方法の一つは，さっきやって見せたように，パーセンタイルで表現することです。しかし問題があるのは，この方法だと "トップが寂しい" ということです。私の友人が集めたサンプルが 1000 人に過ぎなかったとしましょう (これでもまだ新しい質問紙を検証するためには大きいサンプルですが)。そして今回，平均が 50 問中 16 点で標準偏差が 5 だったとします。問題は，このサンプルでは私と同じぐらいの不機嫌度を持っている人が一人もいないということです。

しかし，全てが失われたわけではありません。もう一つのアプローチとして，私の不機嫌さスコアを**標準スコア**に変換するのです。これは z- スコアとも言われています。標準スコアは私の不機嫌さスコアが平均から標準偏差いくつ分上にあるかを表すのです。これを "数学っぽく" いうと，標準偏差は次のように計算できます。：

$$\text{標準スコア} = \frac{\text{ロースコア} - \text{平均}}{\text{標準偏差}}$$

実際数学的には，z スコアについての式は次のようになります。

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

不機嫌さデータに戻っていうと，ダニーの生の不機嫌さデータを標準化された不機嫌さスコアに変換することができます。

$$z = \frac{35 - 17}{5} = 3.6$$

この値を解釈するときに，セクション **??**で触れた，平均から 3 標準偏差範囲にだいたい 99.7% が

入るという，概算を思い出してください。ですから，私の不機嫌さスコアを z スコアにして 3.6 になったということは，実際私はかなり不機嫌状態にあるということです。実際この推論からいくと，私は全体の 99.98% の人よりも不愉快なのです。そうですよね。

ロースコアをより大きな母集団に広げて解釈することを許すとするなら (そしてそれによって任意の尺度の変数を意味のあるものにするなら)，標準スコアは第二の便利な機能を持っていると言えます。標準スコアはロースコアができないような状況でも互いに比較することができます。たとえば，私の友人が 24 項目からなる外向性を測る別の質問紙を持っていたとしましょう。この尺度が全体的に，平均が 13 で標準偏差 4 であり，私のスコアが 2 だたっとしましょう。想像の通り，私の外向性のロースコア 2 を，不機嫌さ質問紙のロースコア 35 と比較するのは意味がありません。この二つの変数のロースコアは基本的に違うもので，いわばりんごとオレンジを比較するようなものです。

標準スコアではどうでしょう？ これはちょっと事情が違います。標準スコアを計算すると，不機嫌さは $z = (35 - 17)/5 = 3.6$，外向性は $z = (2 - 13)/4 = -2.75$ となります。この二つの数字は相互に比較することが できます[*5]私はほとんどの人の中では外向性が低く ($z = -2.75$) ，不機嫌さが高い ($z = 3.6$) のです。しかし私のハズレ具合は外向性よりも不機嫌さの方が大きいといえます。3.6 が 2.75 よりも大きな数字だからです。それぞれの標準化スコアはその観測値が *その母集団においてどのあたりに*落ちるのかを示すので，全く異なる変数についても標準スコア同士を比較することができるのです。

1.6 要約

基礎統計量を計算することは，あなたが実際にデータを取ったとき真っ先にするべきことの一つであり，記述統計量は推測統計よりも単純で理解しやすいので，他の統計の教科書と同じように私も記述統計から説明しました。この章では，以下のトピックスについて議論しました。

- *中心化傾向の指標* 一般的に，中心化傾向はデータがどのあたりにあるのかを教えてくれます。典型的に報告される指標は次の三つでしょう；平均値，中央値，最頻値です (セクション Section **??**)。
- *変動の指標* それに対して，変動の指標はデータがどのように “散らばっているか” を教えてくれます。鍵になる指標としては，次のものがあるでしょう；範囲，標準偏差，四分位範囲です (セクション **??**)。

[*5] いくつかの注意は必要です。変数 A についての 1 標準偏差が，変数 B の 1 標準偏差と “ある意味” 対応しているとは言えないからです。二つの変数に関する z スコアが意味のある比較ができるかどうかを決めるには，常識をはたらかせねばなりません。

- *歪みと尖りの指標* 変数の分布が非対称さの指標 (歪度) と，尖り具合 (尖度) もみてきました (セクション **??**)。
- *JASP で群ごとに変数の要約をする* この本では JASP でデータ分析をすることに焦点化していますから，異なるサブグループそれぞれについて記述統計量を計算するにはどうするかについても触れました (セクション **??**)。
- *標準化スコア* z-スコアはちょっと変わった野獣です。これは記述統計量とはちょっと違いますし，推測統計の話でもありません。これについてはセクション **??**で触れました。この章も理解してもらえたと思います。また後で触れることになります。

次の章では，どうやって絵を描くのかについての話題に移りたいと思います! 誰だって可愛い絵が好きですもんね? しかしそれの前に，重要な点を抑えておきたいと思います。統計の伝統的な入門コースは，記述統計について小さな配分しかせず，1,2 回授業で触れる程度です。授業時間のほとんどの時間は，推測統計学に使われます。というのも，そこが本当に大変なところだからです。それはそれで意味があるのですが，良い記述統計量を選択するという，日々の重要な実践を覆い隠してしまいます。このことを覚えておいて欲しいのです…

1.6.1 エピローグ: 良い記述統計量とは記述的である!

> *一人の死は悲劇である。*
> *数百万の死は統計である。*
> – Josef Stalin, Potsdam 1945

> *950,000 – 1,200,000*
> – ソ連における弾圧の死者数,
> 1937-1938 (**Ellman2002**)

スターリンの悪名高き，数百万人の死に関する統計の特性についての引用は，少し考えてみる必要があります。彼の主張意図は明らかに，個々人の死は我々の心に触れ，無視することはできないけれども，非常に多くの死については理解できないし，結果的に単なる統計であって，無視してしまうことも簡単である，というところにあります。スターリンは，半分は正しいと思います。統計というのは抽象化であり，個々人の経験を超えた出来事の記述であり，可視化されにくいものです。百万人の死が “本当に” どういうことなのかを想像できる人はほとんどいませんが，一人の死は簡単に想像できることができますし，孤独な死は悲劇の感情を呼び起こし，Ellman の冷たい統計的記述の感覚が失われたように感じます。

これはそんなに簡単な話ではないのです。数字がなければ，数えなければ，何が起こったのかの記述がなければ，われわれは本当に何が起こったのかを理解する *機会*すらもてず，この失われた感覚

を呼び起こす機会さえ持つことができません。そして実際には，私はこれを気持ちの良い土曜日の朝に腰掛けながら書いており，世界の半分そしてこれまでの人生でずっと，ソ連の強制収容所から離れたところにいるのですが，Ellman の推定値とスターリンの引用を書く時には鈍い恐怖がズッシリ胃にきて，寒気を覚えます。スターリン主義の弾圧は私の経験を超えたところにありますが，統計データと結びつき，そこに記録された個人史を思うと，私の理解を完全に超えているとはいえません。なぜなら，Ellman の数字は私たちに教えてくれるからです。2 年以上のスターリンの弾圧によって，私の住んでいる街に今生きている全ての男性，女性，子供たちと同じ数の人が消え去ってしまったのだ，ということを。この死の一つ一つに，独自の物語があって，それぞれの悲劇があって，その幾らかは私たちにも知られています。ですから，注意深く選ばれた統計量を見ながら，残虐行為のスケールに焦点化していきましょう。

統計家と科学者の最初の仕事である，データを集めて要約し，何が起こったのかを聴衆に知らせる数字を見つけてくるというのは，簡単なことではないのです。これは記述統計の仕事ですが，数字だけを使って何が言えるかはその仕事ではありません。あなたはデータアナリストであり，統計ソフトパッケージではないのです。あなたがすべきはこれらの *統計量*を取り出して，*記述*に持っていくことなのです。あなたがデータを分析するとき，数字のコレクションをリストアップするだけでは十分ではありません。忘れてはいけないのは，あなたは人間の，聴衆を相手にコミュニケーションしようとしているということです。数字は重要ですが，あなたの聴衆が理解できるような意味のあるストーリーと一緒でなければなりません。あなたはフレーミングについて考える必要がある，ということです。文脈について考えなければなりません。あなたの統計量が要約した，一つ一つの出来事について考えなければなりません。