いのち嬉しき解除の御沙汰　＜t-検定ドリルでよくある間違い＞

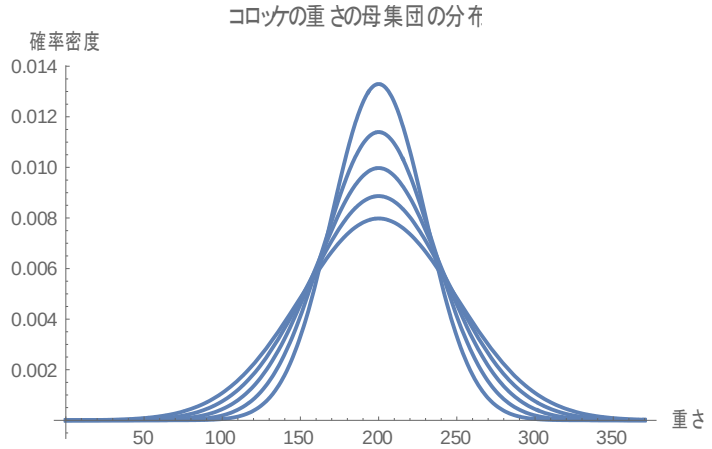
2021 年 9 月下旬　白田由香利

何回目かの非常事態宣言が解除されるとのニュースが走った。これでガラリと世の中が変わるわけでもないが、とても嬉しい。ファミリーの雑事の諸々もあり、義理の妹氏と近所の紅茶専門店で待ち合わせ、近況を報告し合った。二人とも少しおめかしして、私はお着物。このネコ柄の帯は、コロナ直前のクリスマスイブに家族 3 人で銀座でわいわい選びながら買ったものだ。混雑するパン屋の木村屋でやっと席を見つけ、息子はアンパンをむしゃむしゃ食べていたなあ、と思い出す。モノには、楽しい記憶が染みついていて、その記憶が辛い時を乗り切るパワーを与えてくれる。人間、日常の有難さは、奪われてみて初めて分かる。当たり前に買い物をすませ、友人とお茶をして、外食を楽しむ。大学に行き、ＰＣを立ち上げパワーポイントをスクリーンに映し、講義をし、学生の質問を対面で受ける。インドネシアに行って研究成果報告をしあい、次回の研究プロジェクトの相談を対面でする。そうした普通の日々よ早く戻って来てくれ、と切に願う。大事な日常生活は無くして初めて痛感する。以下の NHK のサイトにコロナ関連のニュースがまとめてあった。これは貴重なデータベースなので皆様もご活用ください。
https://www3.nhk.or.jp/news/special/coronavirus/chronology/?mode=digest&target=202110

さて、統計的仮説検定 t-検定を使った仮説検定でよくある間違いを説明していく。問題として、コロッケ屋さんのコロッケの重さを使う。
＜問題＞　お肉屋さんがコロッケ 200 g と言って売っている（これが理論上の 仮説)。どうも少ない気がしてならない(仮説からのずれ)。そこで以下のように帰無仮説と対立仮説をたてて検定してみる。

- 帰無仮説：コロッケの母集団の平均は 200 g である　$\mu = 200$
- 対立仮説：コロッケの母集団の平均は 200 g より少ない　$\mu < 200$

さて、初めにすることは、お肉屋さんに行って標本を買ってくることだ。何個買うかで、使う分布が違ってくる。5 個買ってきたとしよう。標本サイズ n=5 である。
母分散が分かっていないときは、正規分布ではなく、t-分布を使う。母平均が分かっても母分散が未知の場合、分布の形状はどうなるか分からない。母平均が分かっても母分散が未知の場合、分布の形状はどうなるか分からない（下図参照）。



お肉屋さんがコロッケを売るとき、重さの平均は言ってくれるかもしれないが、重さの分散を言ってくれることはまずない。よって、この場合は t-分布を使う t 検定となる。

以下のようにその標本に対する統計量 t 値を計算する。

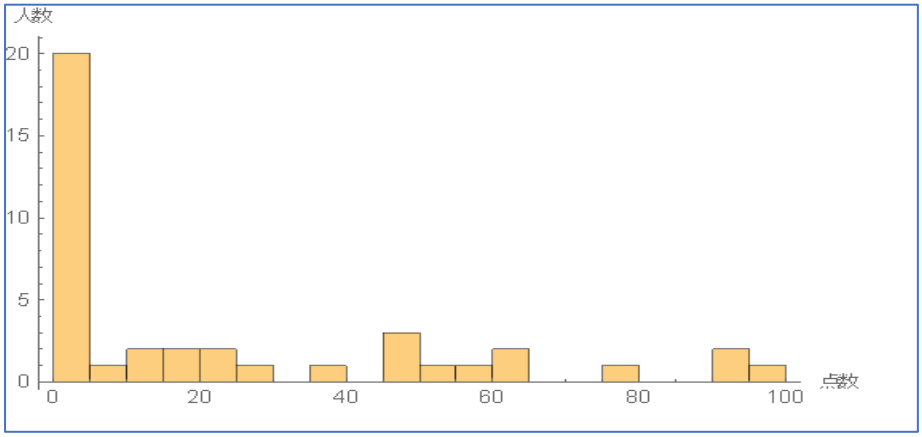
$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}$$

母分散が分からないので、代わりに標本分散（不偏分散）のs^2を使うところに注意する。標本分散は(n-1)で割ることに注意する。

次に注意すべきは、買うコロッケはランダム抽出される必要があることである。まとめて 5 個買うのではなく、理想的には日時を変えるなどして、ランダムに標本抽出をする必要がある。

さらに、n=5 のように 30 よりも標本数が小さい場合、1 個の重さの母集団は正規分布に従っていることを確認せよ。

コロッケの 1 個の重さは正規分布に従うので問題ない。しかし、世の中には正規分布に従わない確率変数もある。例えば、難しい数学の試験の点の分布など、半数近くが 0 点で点を取った人がまばらに分布しているような分布をとる（もちろん、それではきちんと成績評価できないので、そういう問題は不適切であると言えるが）。その場合は、点数の分布は正規分布にはならない。



t -検定をする前に、この重さ(確率変数)の分布は正規分布かしら？と気にしなくてはいけない。

まとめると、母分散が分からない場合で、t-検定をしてもいいのは、以下のどちらかの条件を満たす場合である。

（1）n が 30 未満であれば、母集団は正規分布に従っていなくてはならない。

（2）n が 30 以上であること。(母集団はどのような分布でも構わない)

（2）の場合、どうして母集団はどのような分布でも構わないのであろうか？

答えは、標本サイズ n が 30 以上のときは、中心極限定理から、標本平均の分布は正規分布に近づくからである。母集団が正規分布でなくても、標本平均の分布は正規分布になる。中心極限定理から、標本平均の正規分布の分散は n 分の 1 になる。だから、標本分散を n で割って、そのルートをとった以下の式が、標本平均の分布の標準偏差となる。

$$\sqrt{\frac{s^2}{n}}$$

そして　$\bar{X}-\mu$　（標本平均の母平均からのずれ）を、「標本平均の分布の標準偏差」で割った値がt-値である。

では、最後にまとめの質問。

Q１：$\sqrt{\frac{s^2}{n}}$　のように 1/n で割っているのは、どの理論に由来してですか？

Q２：標本分散s^2を計算するとき、偏差の平方和を(n -1)で割るのはなぜですか？

分からなかった人は、前の統計についての説明を読み返してみてください。

終わり

（答え）

Ａ１：中心極限定理

Ａ２：自由度が１減っているから。その理由は、母平均μの代わりに標本平均$\bar{X}$を使ったので、鎖につながれたようになって、最後の１個のデータの自由度がなくなるので。

引用元：いのち嬉しき撰集のさた　　向井去來