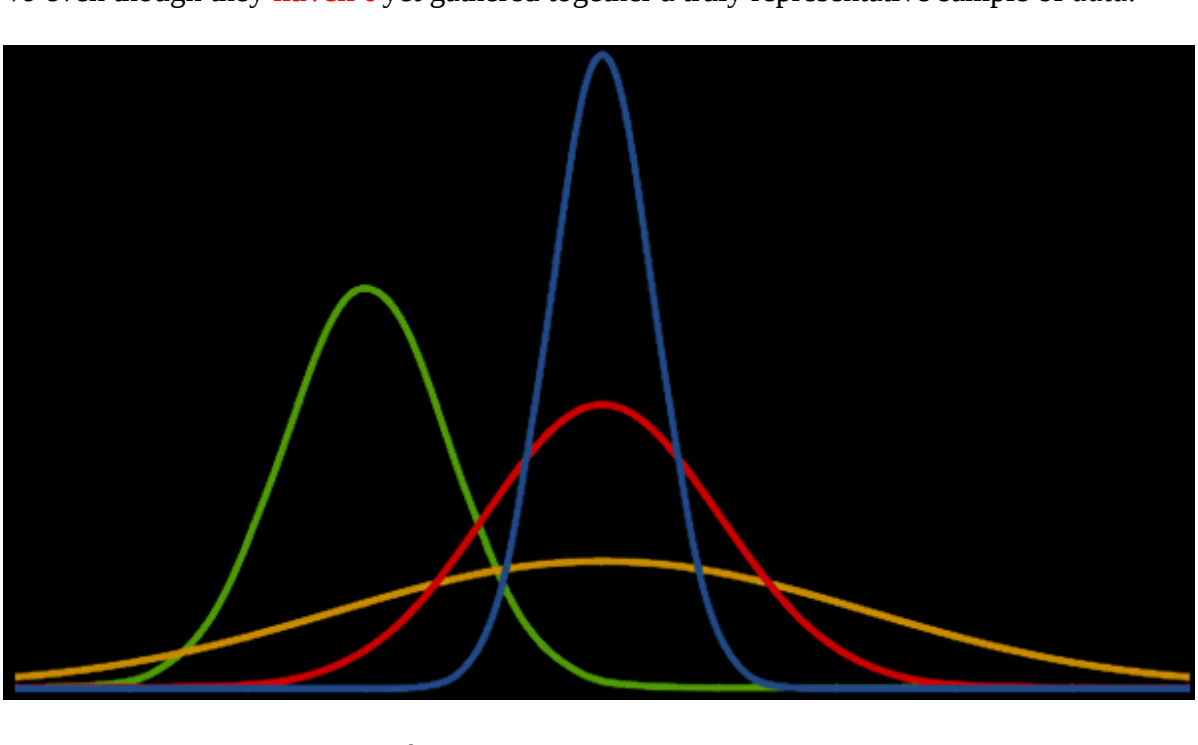
23rd May 2024

Many people assume the natural distribution of their *data* will eventually resemble some form of *bell curve* even though they **haven't** yet gathered together a truly representative sample of *data*.

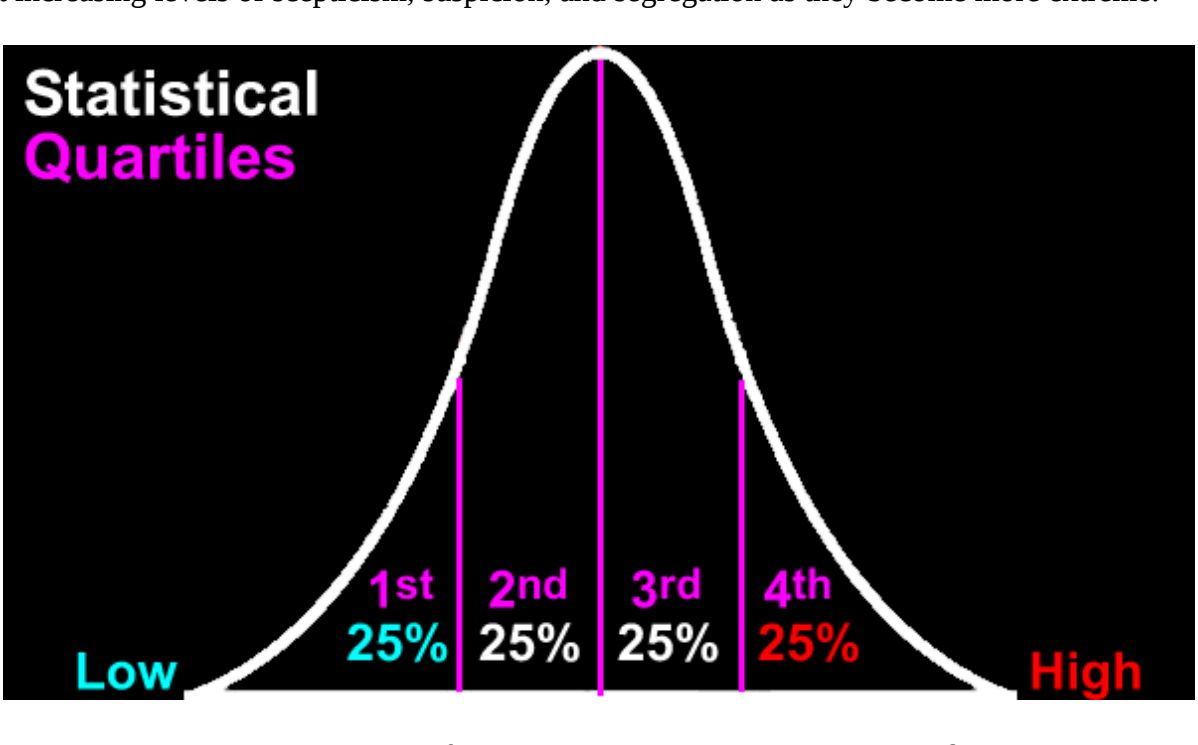


> **Normal distributions** … are often **used in the natural** and social **sciences to represent** real-valued **random variables whose distributions are not known**.
>
> Wikipedia - Normal Distribution
> https://en.wikipedia.org/wiki/Standard_distribution

Individual interpretations of *bell curves* tends towards considering the ***middle* 50%** of the data as the ***normal range*** of values.

The [unmentioned] implication being:

The ***low end* 25%** and ***high end* 25%** of samples are regarded as ***abnormal ranges*** of values that warrant increasing levels of scepticism, suspicion, and segregation as they become more extreme.

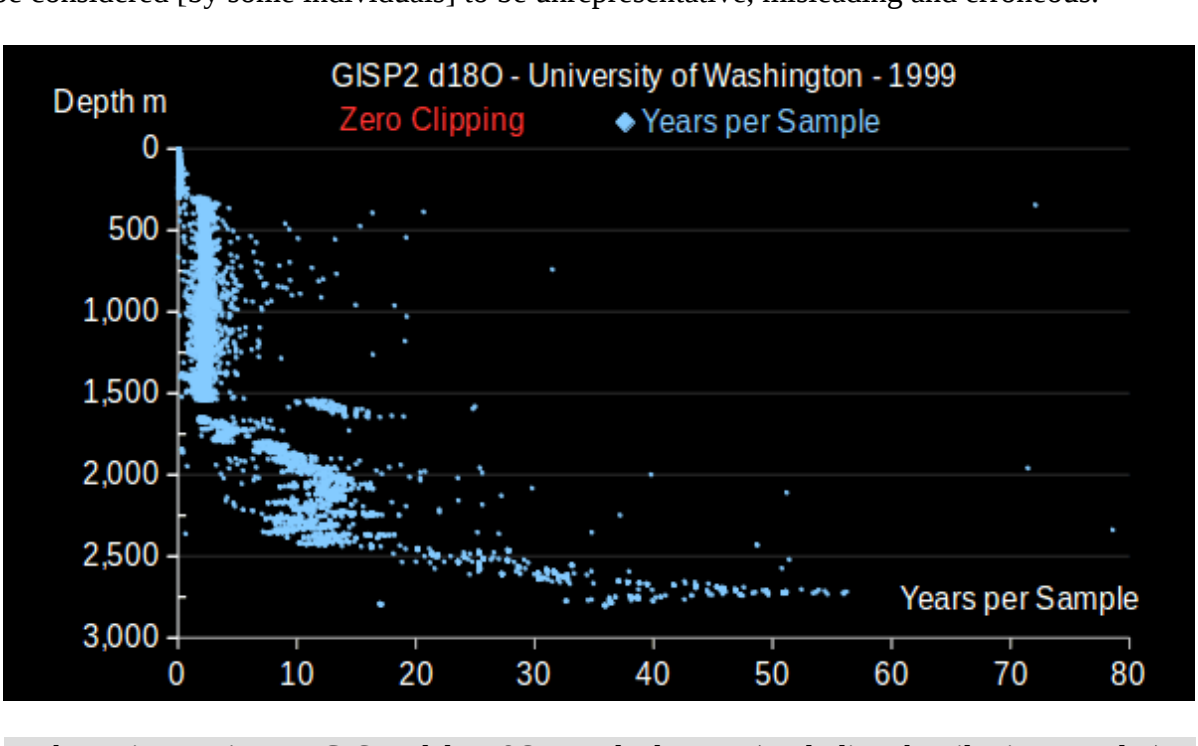


> In statistics, **quartiles** are a type of quantiles which **divide the number of data points into four parts**, or quarters, **of more-or-less equal size**. The data must be ordered from smallest to largest to compute quartiles; as such, quartiles are a form of order statistic.
>
> Wikipedia - Quartile
> https://en.wikipedia.org/wiki/Quartile

This leads some people who are in the business of managing appearances [aka expectations aka opinions aka perspectives aka beliefs] to think they should [for either **good** or **nefarious** purposes] ***fiddle about*** with the more extreme *outlier* data.

For example:

The outlier *Years per Sample* data points derived from the published **GISP2 Ice Core** chronology **could** be considered [by some individuals] to be unrepresentative, misleading and erroneous.

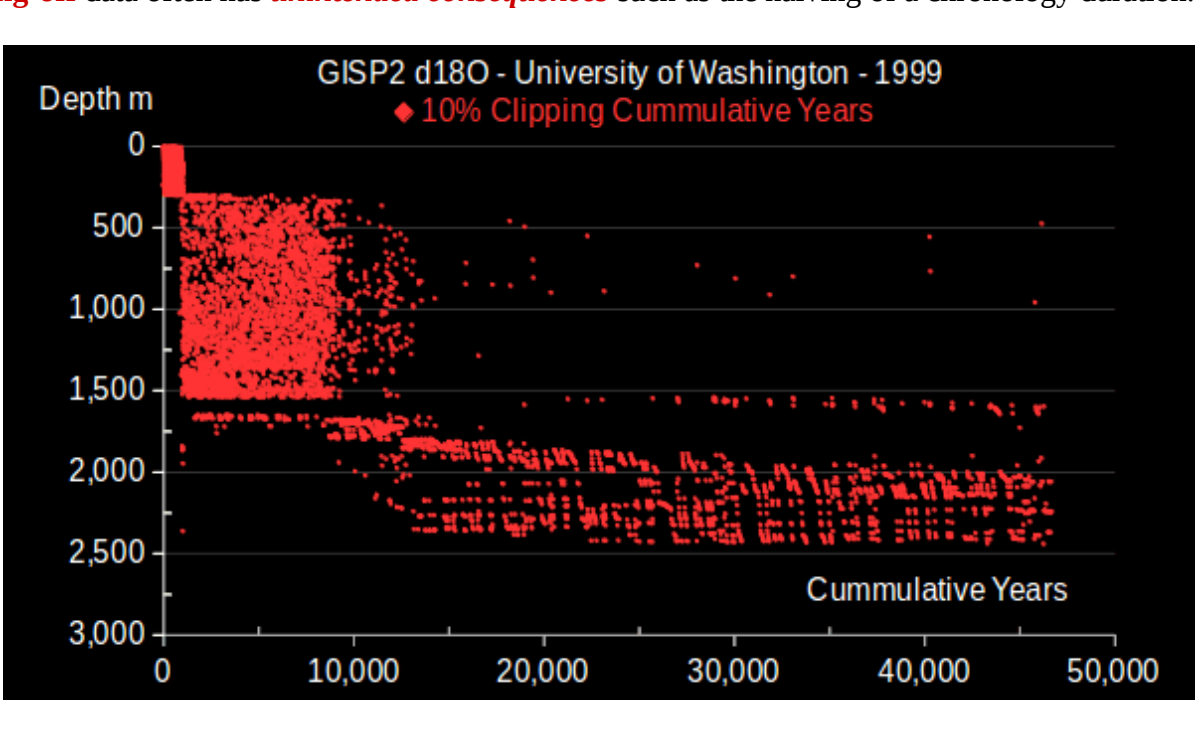


> The entire continuous GISP2 delta 18O sample data set (excluding the silty ice samples)
> University of Washington's Quaternary Isotope Laboratory - 5 March 1999
> https://web.archive.org/web/20060517193048/http://depts.washington.edu/qil/datasets/gisp2_measured.txt

These individuals **could** consider **clipping-off** [like hedge trimmings] a percentage of the data values that ***stick out*** in the ***low-end*** **and *high end*** of the dataset.

However:

Clipping-off data often has ***unintended consequences*** such as the halving of a chronology duration.



Evidently:

A more sophisticated technique is required when you're ***fiddling about*** with data.

> While cleaning strategies remove glitches from data, they do not necessarily make the data more usable or useful … The data can be changed to such an extent that they no longer represent the underlying process that generated the data.
>
> Statistical Distortion: Consequences of Data Cleaning
> Tamraparni Dasu and Ji Meng Loh
> Proceedings of the VLDB Endowment - August 2012
> https://www.researchgate.net/publication/230639505

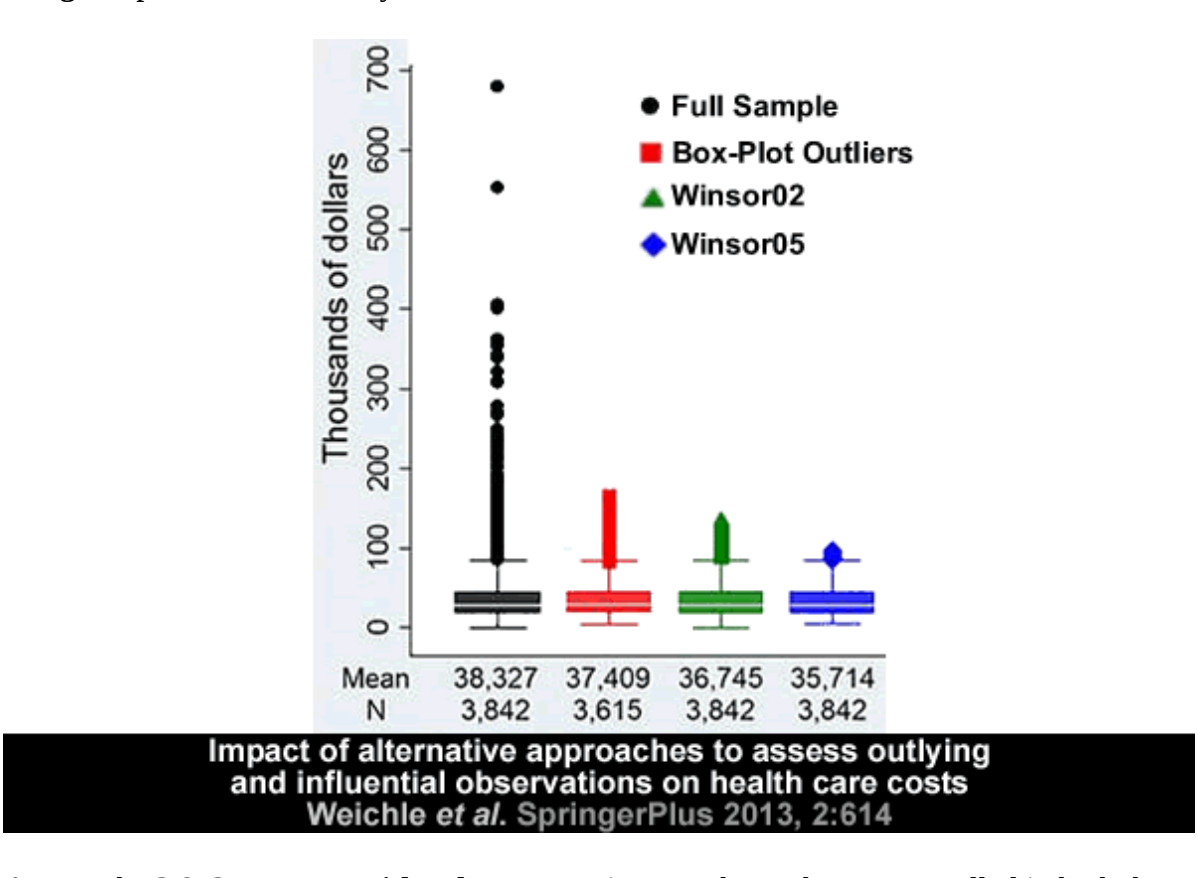
A more sophisticated technique like **Winsorizing**.

> **Processing**:
> Dark frame calibration, Star Alignment, **Winsorized** Sigma Average Stacking + Drizzle, Starmask and Morphological Transformation, Histogram Transformation in PixInsight. Combined as Ha-OHS in Photoshop, Inverted SCNR + regular SCNR in PixInsight, Levels and Curves in Photoshop with inverted layer masking + final resize + Topaz noise reduction.
>
> The Sulphur in NGC 3603 and NGC 3576 - Dylan O'Donnell
> https://deography.com/the-sulphur-in-ngc3606/

> The technique of **Winsorizing** is used to estimate the population mean (and variance), to form confidence intervals, and to test hypotheses about the population mean in situations where outliers are suspected (Dixon and Tukey, 1968).
>
> Statistical Analysis: A Computer Oriented Approach - 1979
> A A Afifi and S P Azen
> https://archive.org/details/statisticalanaly0002afif/page/114/mode/1up

> **Winsorizing** or winsorization is the transformation of statistics by **limiting extreme values** in the statistical data to reduce the effect of possibly spurious outliers. It is named after the engineer-turned-biostatistician **Charles P. Winsor** (1895-1951).
>
> The effect is the same as clipping in signal processing.
>
> **The distribution** of many statistics can be **heavily influenced by outliers**.
>
> A typical strategy is to set all outliers to a specified percentile of the data; for example, a **90% winsorization** would see
>
> all data below the 5th percentile **set** to the 5th percentile,
> and data above the 95th percentile **set** to the 95th percentile.
>
> Winsorized estimators are usually **more robust to outliers** than their more standard forms, although there are alternatives, such as **trimming**, that **will achieve a similar effect**.
>
> Wikipedia - Winsorizing
> https://en.wikipedia.org/wiki/Winsorizing

Winsorizing can produce some really remarkable results.

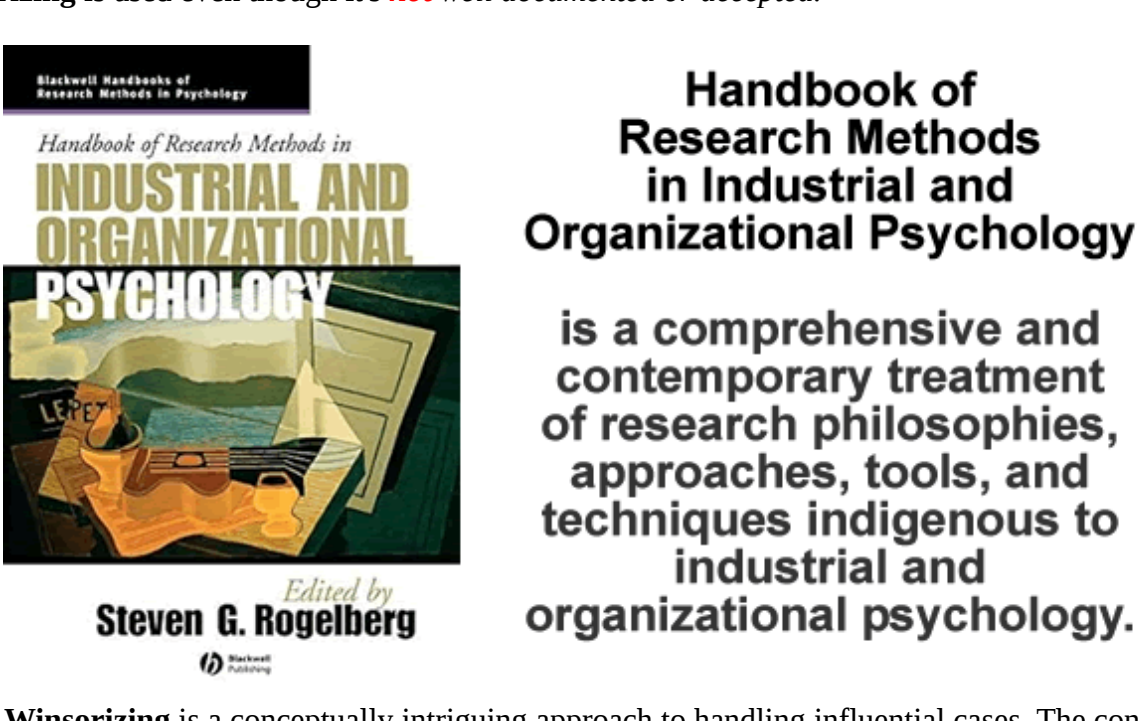


> Among the **3,842 veterans with colon cancer** in our cohort who were enrolled in both the VA and Medicare between 1999 and 2001 … The **average cost** of colon cancer episodes for the cohort was **$38,327** … with a range of **$43 to $679,472**.
> ...
> **The box-plot method identified 227 observations as outlying**. Based on their distribution, 45 observations were upper outlying values and 182 were lower outlying values.
> ...
> **Winsorization at the 2nd and 98th percentiles replaced 152 observations**
> (76 observations in the lower end and 76 in the upper end). …
> Winsorization at this level replaced 2% of the skewed observations to the right.
> ...
> **Winsorization at the 5th and 95th percentiles replaced 384 observations**
> (192 observations in the lower end and 192 in the upper end).
> Winsorization at this level replaced 5% of the skewed observations to the right
>
> Impact of alternative approaches to assess outlying
> and influential observations on health care costs
> Thomas Weichle, Denise M Hynes, Ramon Durazo-Arvizu,
> Elizabeth Tarlov, Qiuying Zhang
> November 2013 - SpringerPlus 2(1):614
> https://www.researchgate.net/publication/259156208

> In descriptive statistics, a **box plot** or boxplot is a method for **graphically demonstrating** the locality, spread and skewness groups of **numerical data through their quartiles**.
>
> In addition to the box on a box plot, there can be **lines** (which are called whiskers) extending from the box **indicating variability outside the upper and lower quartiles**, thus, the plot is also called the box-and-whisker plot and the **box-and-whisker diagram**.
>
> Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot.
>
> Wikipedia - Box Plot
> https://en.wikipedia.org/wiki/Box_plot

And

Winsorizing is used even though it's ***not*** *well documented or accepted.*

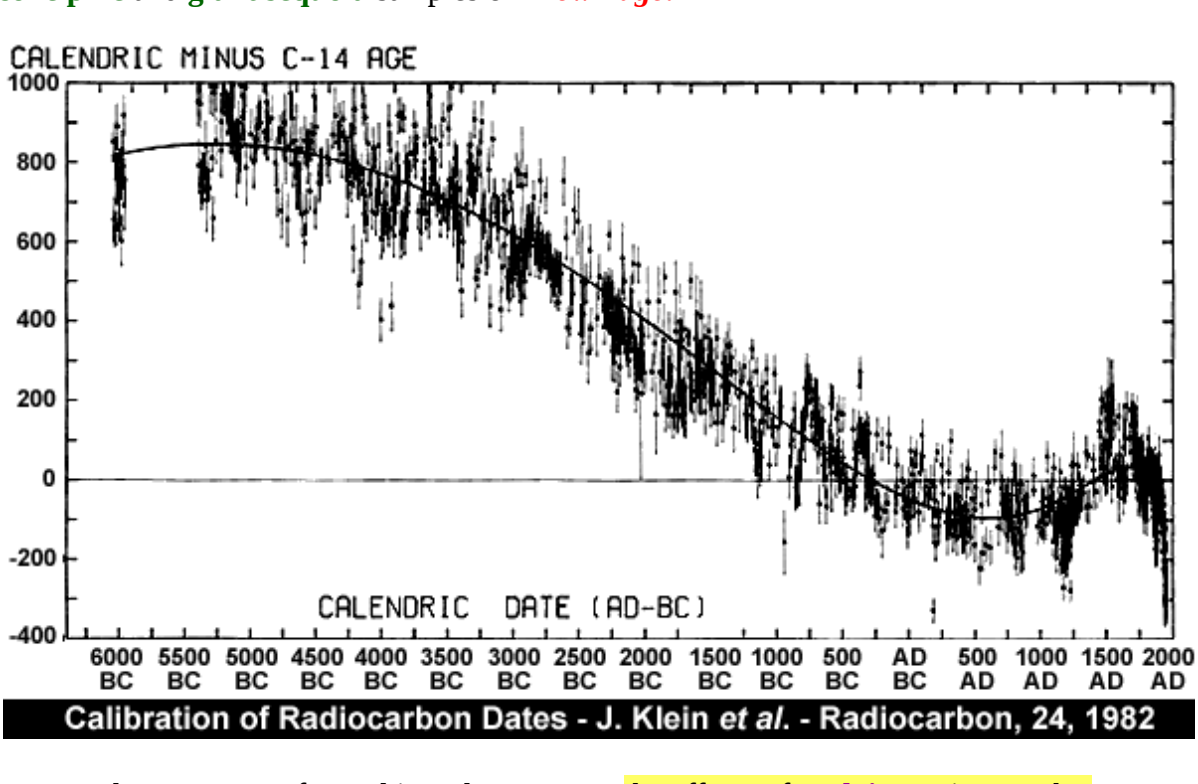


> **Winsorizing** is a conceptually intriguing approach to handling influential cases. The concept underlying Winsorizing is not to delete the case in question, but to modify its score so it is no longer deviant from other cases. Early research analyzed univariate statistics such as the mean (Guttman and Smith, 1969) and the standard deviation for fairly small samples (Guttman and Smith, 1971). There were a variety of mathematical rules to adjust deviant scores (e.g., Guttman, 1973). For example, one rule was to set the deviant number to the next nearest number in the data set.
>
> The advantages of Winsorizing parallel the advantages of various missing data techniques. The approach keeps, and does not modify, other scores in a case when a univariate approach is used and it preserves all this information from possible deletion. The potential disadvantages include the difficulty of determining the bivariate and multivariate statistical space and how to modify cases in such a way as to change values and preserve as much original data as possible. The approach is also not well documented or accepted at this time.
>
> Handbook of research methods in industrial and organizational psychology
> Editor: Steven G Rogelburg - 2002
> Amazon US: https://www.amazon.com/dp/B000WNFPA0
> Amazon UK: https://www.amazon.co.uk/dp/B000WNFPA0

> Upon closer inspection of the data, we observed an outlier score of 158.40, which was three standard deviations below the mean language arts scale score in the sample. In this case we used the **Winsorizing** procedure to substitute the outlier score with a score of 173, which was one standard deviation higher (Field, 2013). A score of 173 was just 2 points away from the next highest value, 175, which was not an outlier in our data set.
>
> Education Policy Perils: Tackling the Tough Issues
> Editors: Christopher H Tienken and Carol A Mullen - 2015
> Amazon US: https://www.amazon.com/dp/B0BL5DX44C
> Amazon UK: https://www.amazon.co.uk/dp/B0BL5DX44C

> A Winsorization adjustment was applied to four district outlier weights. …
> A Winsorization adjustment was applied to four outliers. …
> A Winsorization adjustment was made for seven extreme school weights. …
> A Winsorization adjustment was made for seven extreme weights. …
> A Winsorization adjustment was applied to seven extreme school weights …
>
> State and Local Implementation of the "No Child Left Behind Act"
> Volume III--Accountability under "NCLB" Interim Report
> U.S. Department of Education - 2007
> https://archive.org/details/ERIC_ED499023/page/n155/mode/1up

Winsorization was employed in the construction of the **1982 radiocarbon** chronology based upon **bristlecone pine** and **giant sequoia** samples of ***known age***.



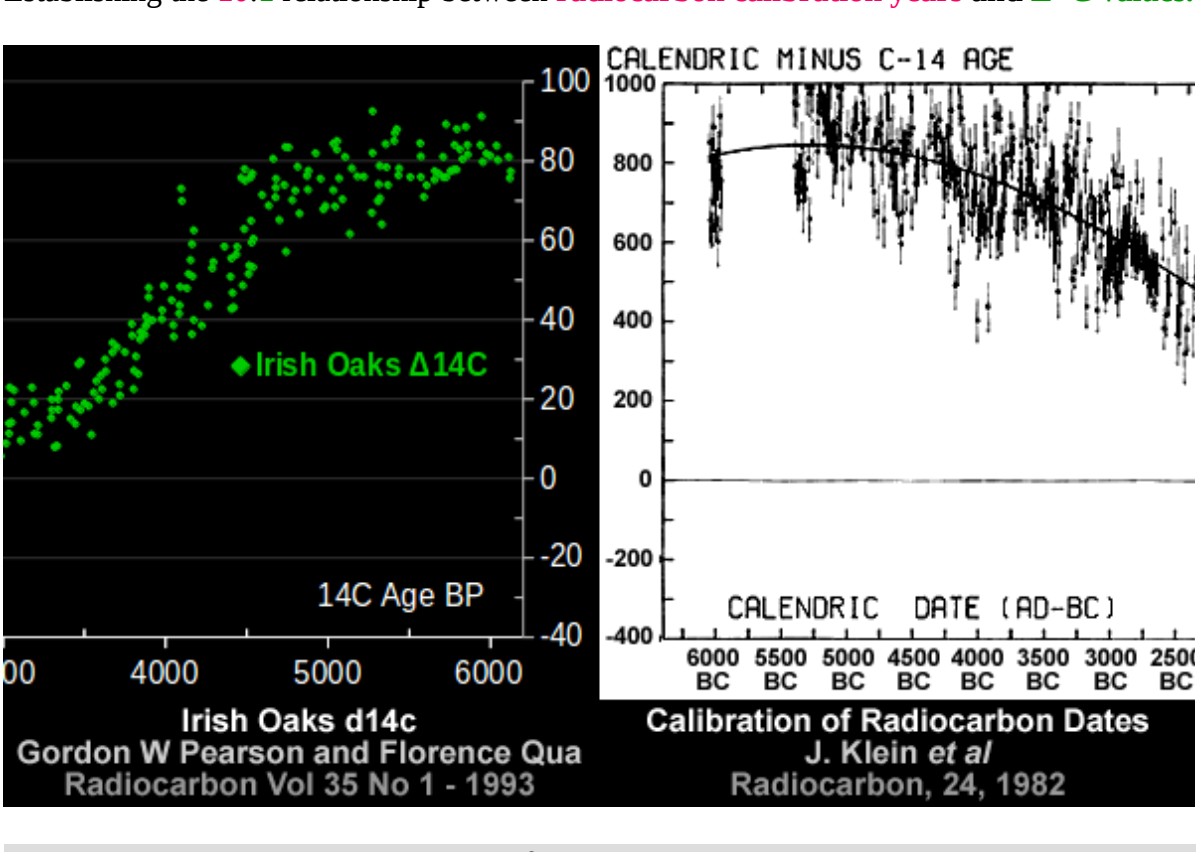
> Two analyses were performed in order to assess the effects of **outlying** points on the **calibration** function.
>
> The first analysis used the unmodified data base as described in the section on data, whereas the second analysis used a "**winsorized**" data set in which the residuals used for **winsorization** were taken with respect to the function calculated on the unmodified data.
>
> "**Winsorization**" is a process which reduces the effect of a few **aberrant measurements** by limiting the effect on the mean of a single **outlying** point to less than ~2.56s/n, where s is the standard error estimated from the fourth quintile of the variance of the data, and n is the number of points in the interval. **Winsorization**, as employed here, is described elsewhere (Dixon, 1960).
>
> **Winsorization** was used **instead of** a **simple rejection** of "**outlying**" points …
>
> **Calibration of Radiocarbon Dates**
> Tables based on the consensus data of the
> Workshop on Calibrating the Radiocarbon Time Scale
> Jeffrey Klein, J C Lerman, P E Damon, and E K Ralph
> Radiocarbon - Volume 24 - Number 2 - 1982
> https://journals.uair.arizona.edu/index.php/radiocarbon/article/download/748/753

> A censored sample is a sample lacking one or more observations at either or both extremes with the number and positions of the missing observations known.
>
> Censoring may take place naturally i.e., an observation has a magnitude known only to be more extreme than the other observations in the sample.
>
> Censoring may also be imposed by the experimenter who from past experience knows that extreme observations are so unreliable that their magnitudes should not be used as observed.
> …
> **Winsor** and perhaps others have suggested using for the magnitude of an extreme, poorly known, or unknown observation the magnitude of the next largest (or smallest) observation.
> …
> For non-symmetrical censoring, it is demonstrated that optimum simple estimators of the mean result from these "**Winsorized**" estimators.
>
> Simplified Estimation From Censored Normal Samples - W J Dixon
> The Annals of Mathematical Statistics- Volume 31 Issue 2 - June 1960
> https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-31/issue-2/

Therefore:

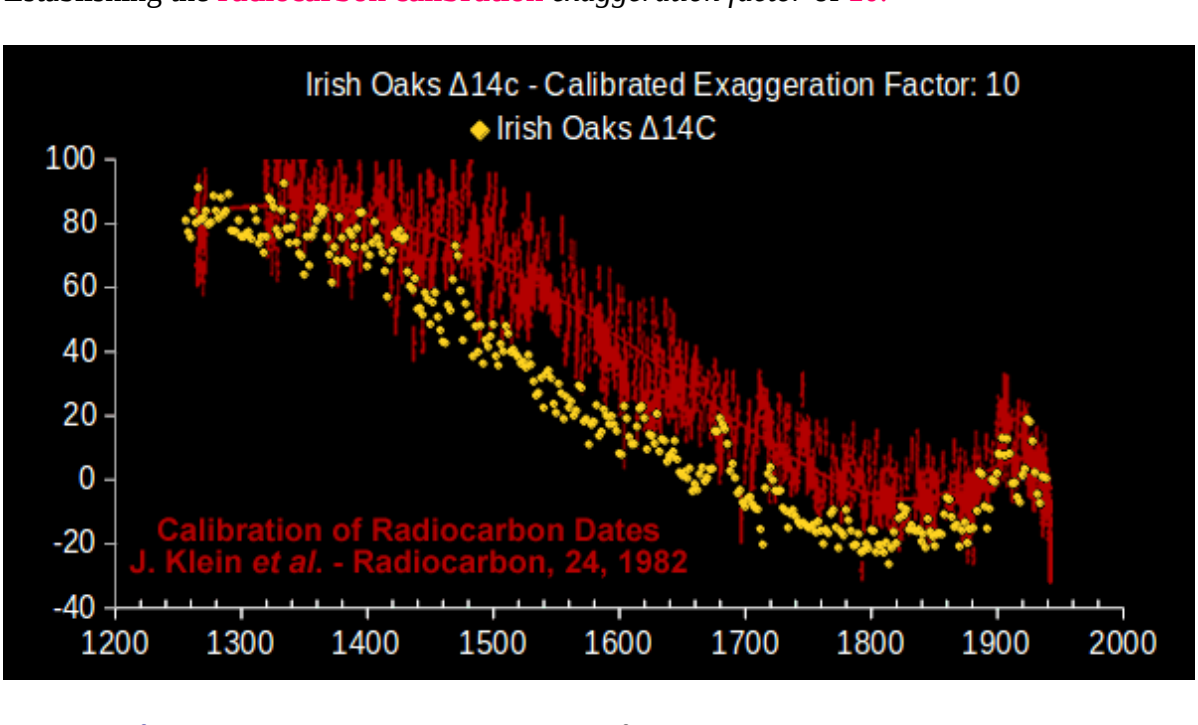
Winsorization was involved in:

► Establishing the **10**:**1** relationship between **radiocarbon calibration years and $\Delta^{14}C$ values.**



> High-Precision 14C Measurement of Irish Oaks to Show the Natural 14C Variations
> Gordon W. Pearson and Florence Qua - Radiocarbon, Volume 35, No. 1, 1993
> https://journals.uair.arizona.edu/index.php/radiocarbon/article/download/1556/1560

► Establishing the **radiocarbon calibration** *exaggeration factor* of **10**.



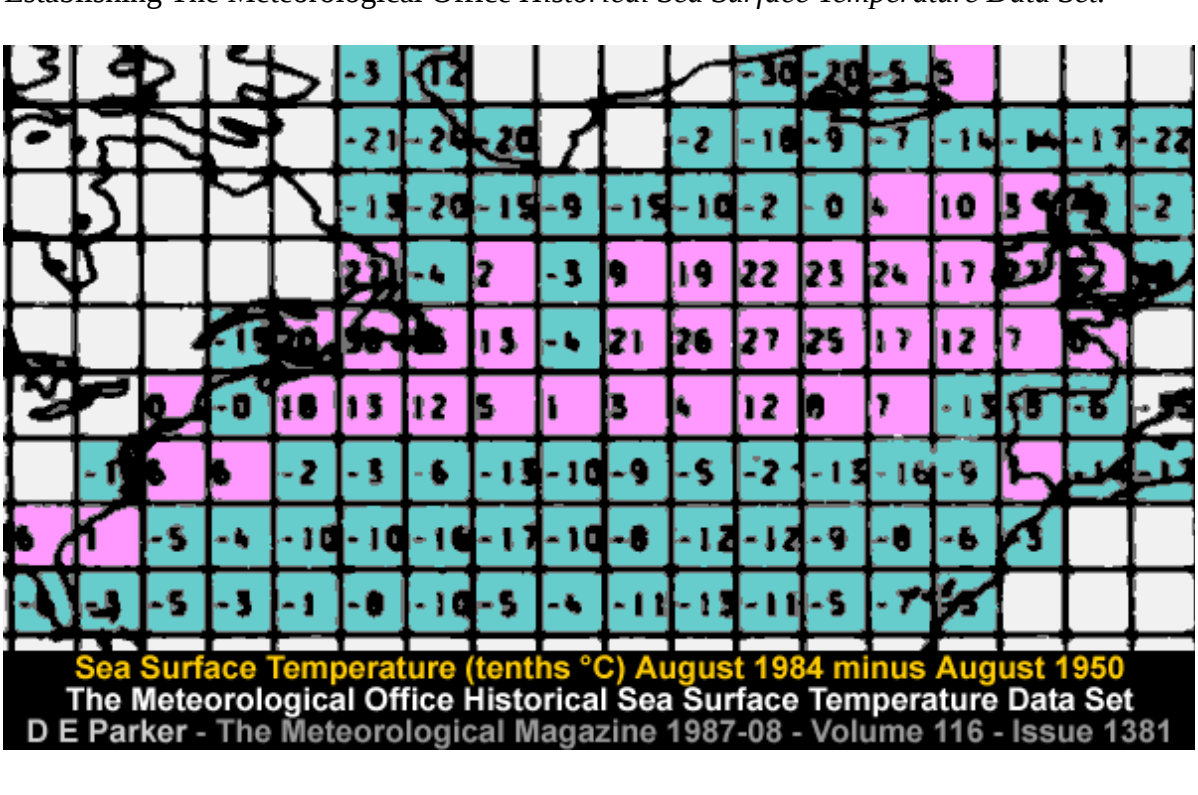
The rule of thumb is to divide the ***ancient age*** of a **bristlecone pine** or **giant sequoia** by **10**.

> Malaga Bay - Methuselah and Mini-Me
> https://malagabay.wordpress.com/2024/05/16/methuselah-and-mini-me/

And

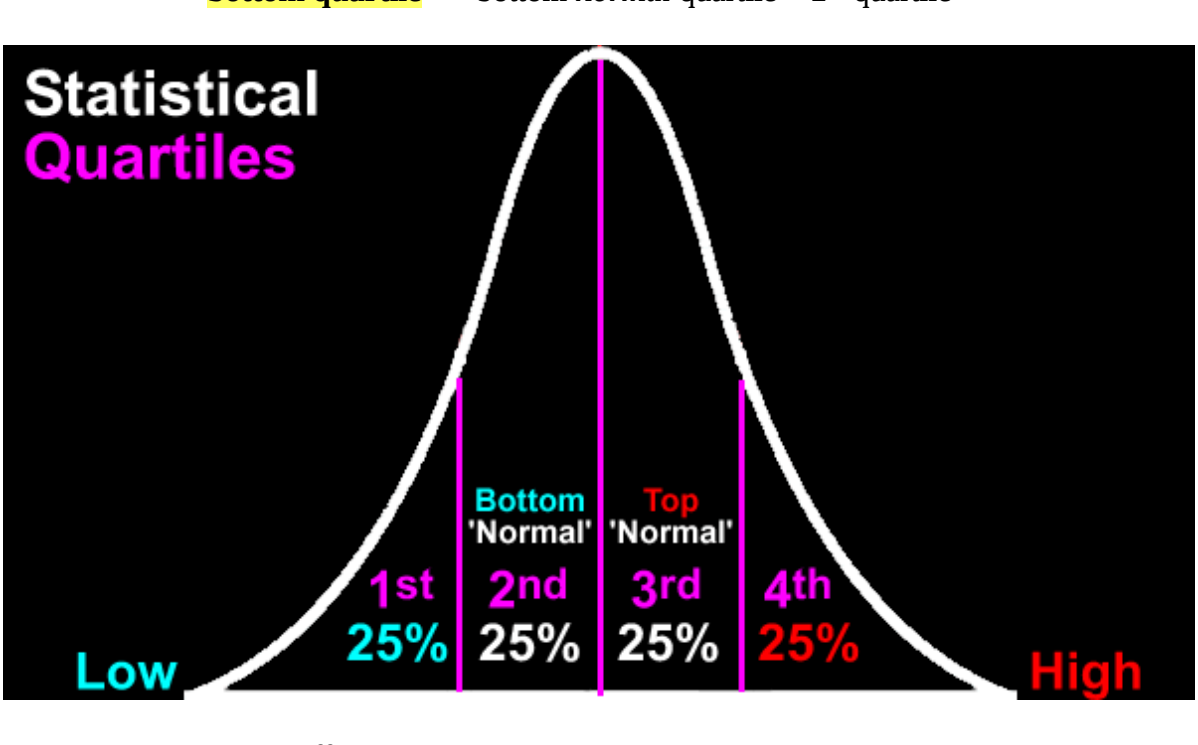
Winsorization was involved in:

► Establishing The Meteorological Office *Historical Sea Surface Temperature Data Set.*



> **Quality control**
> Observations of SST are beset by systematic biases, individual inaccuracies, and irregular distribution in space and time.
>
> Systematic biases occur because of changes in instrumentation, siting, or procedures.
>
> The most notable example is the change from uninsulated bucket measurements (which were taken until the Second World War) to a mixture of engine-intake, hull-sensor, or insulated bucket readings.
> …
> Individual inaccuracies were treated in the following way.
>
> First, no SSTs were included in the main marine data bank
> if they were outside the range -5 °C to 35 °C.
>
> Second, a provisional climatology with 1° latitude X 1° longitude and 5-day resolution was formed during the creation of MOHSST, and all SSTs deviating from this climatology by more than 6 °C were excluded.
>
> Third, after averaging over 1° latitude X 1° longitude areas for 5-day periods, the SSTs were converted into deviations from the provisional climatology and then subjected to a modified averaging process known as '**Winsorization**' (Afifi and Azen 1979).
>
> In this computation, which was made for each 5° latitude X 5° longitude area and month, values exceeding the **top quartile** were replaced by that quartile, and values below the **bottom quartile** were replaced by the bottom quartile.
>
> The adjusted set of values was then averaged. The resulting average is **less influenced** by **outlying values** than a straightforward average would be.
>
> The Meteorological Office Historical Sea Surface Temperature Data Set
> D E Parker - The Meteorological Magazine 1987-08 - Volume 116 - Issue 1381
> Meteorological Office
> https://archive.org/details/sim_meteorological-magazine_1987-08_116_1381/page/n23/mode/1up

Translation: "**top quartile**" = top ***normal*** quartile = 3rd quartile
"**bottom quartile**" = bottom ***normal*** quartile = 2nd quartile



Whether The Meteorological Office still employs **Winsorization** may [or may **not**] be revealed when they release their ***new version*** of the ***Central England Temperature*** dataset in 2024.

> **New CET Version Release**
> A **new version** of the CET dataset is planned for release in **Spring 2024**.
> This version, and subsequent incremental versions, will form part of a new annual release cycle for the CET dataset in which the previous complete year's series values are **recalculated** using **quality controlled** temperature observations from selected CET stations.
>
> Central England Temperature dataset
> Hadley Centre - The Meteorological Office
> https://www.metoffice.gov.uk/hadobs/hadcet/

As always:

Review the evidence and draw your own conclusions.