Analysis of parallel temperature data using t-tests – a case study

Dr Bill Johnston

www.bomwatch.com.au

Summary

A protocol is suggested whereby before undertaking paired and un-paired t-tests of daily temperatures measured in-parallel by different instruments, properties of datasets are examined and steps taken to mitigate autocorrelation, which is the interdependence of values at one time on observations for previous times. Also, as the significance of test outcomes increases as the numbers of samples increase, an empirical measure of whether a *significant* or *highly significant* difference is meaningful in the overall scheme of things is important. The use of paired verses unpaired t-tests for comparing instruments that cannot measure the same parcels of air, 100% of the time, is also discussed.

While for daily data some autocorrelation may be unavoidable, with its effects much diminished, neither paired nor unpaired t-tests detected significance in the difference between daily maximum temperature measured by thermometers, and by the automatic weather station in 60-litre Stevenson screens located 179m apart at Townsville airport. The cause of autocorrelation in daily maximum temperatures at Townsville is outlined.

1. Background

Recent exchanges at https://wattsupwiththat.com/ have emphasised the need for a protocol-based approach to using t-tests to compare parallel daily temperatures measured by different instruments housed in the same or different Stevenson screens. While the t-test is arguably the most frequently used statistical test it is also frequently misused, mainly by ignoring assumptions on which it is based. The unpaired or 2-sample t-test calculates the *probability* (*P*) that the mean for each instrument or screen *is the same*, while the paired or repeated-measures t-test is a one-sample test of whether the mean of the differences between instruments *is zero*.

A low *P*-value, typically less than 0.05 (5% or 1 in 20), rejects the P_{same} or $P_{diff=0}$ hypothesis in favour of the alternative, which is that instruments/screens/sites are *significantly* different. The words *significant* and *highly significant* indicates group means are unlikely to be the same, or for the paired test, that the mean of their sign-preserved differences is unlikely to be zero. P-levels are usually specified as $P<0.05$ (significant), $P<0.01$ (highly significant), or given as $P = 0.xxx$.

The purpose of this note is to outline in general terms, how t-tests are used and their limitations. The research question is whether Tmax measured by different instruments is the same, or whether the mean of their differences equals zero. Parallel data for Townsville airport is used as the case study.

1.1 Paired or un-paired

As opposed to comparing the means of two independent groups, such as randomly chosen subjects receiving Treatment(a) verses another group receiving Treatment(b), paired t-tests compare data that are in the form of matched-pairs (i.e., each subject delivers data for both conditions). For instance, liveweight of *the same* animals measured before and after a new diet or intervention. Pairing accounts for variation *within subjects given the same treatment*, thus

the paired t-test is more likely to show differences are significant, than if the same data were compared as treatment groups.

In this context, *independent* means that samples or observations are from different populations, whereas *paired* implies a connection between them. Whether two types of unrelated instruments measuring properties of the same airmass on the same day constitute a data-pair is therefore debatable. Nevertheless, there are pitfalls in applying the test to large numbers of closely-spaced timeseries.

1.2 Assumptions are important

Validity of all parametric hypothesis tests depends on underlying assumptions, and for t-tests the most important is that *observations* for one subject, or at one time do not predict observations at subsequent times. *Serial dependence* (also referred to as autocorrelation) inflates the likelihood of *false positives*, i.e., finding differences are significant when they are not. Autocorrelation is common in sequential data and the shorter the time interval, the more likely will successive data be serially correlated, hence the name autocorrelation.

As they are affected by cycles and sequences of dry and wet days etc., researchers must verify that daily temperatures measured by two instruments housed in the same or different Stevenson screens are independent of previous data. Likewise, for the differences between serial data-pairs.

How the t-test works is explained in simple terms here: https://blog.minitab.com/en/statistics-and-quality-data-analysis/what-is-a-t-test-and-why-is-it-like-telling-a-kid-to-clean-up-that-mess-in-the-kitchen and restated, the t-value is the ratio of signal to noise – the strength of the signal (the difference being adjudicated), divided by noise in the data (variation as measured by the standard deviation (pooled for un-paired t-tests) adjusted for the number of samples).

There is a large body of information in the public domain that explains the t-test and its strengths and weaknesses (e.g., https://vasishth.github.io/Freq_CogSci/common-mistakes-involving-the-paired-t-test.html) and practitioners should avoid the pitfalls of claiming differences are *significant*, when the wrong test was used, test assumptions were violated, or differences may not be meaningful or consequential.

1.3 The *size* of a difference verses *significance*

A problem with datasets consisting of more than around 60 observations is that as the numbers of samples increase, trivial differences can become increasingly significant.

As explained here: https://stats.stackexchange.com /questions/4075/how-to-perform-t-test-with-huge-samples standard errors decline as the numbers of samples increase, which markedly deflates the *P*-level of the test (i.e., *P*-values decline therefore become more significant). In their paper: *Too big to fail: large samples and the* p-*value problem* (https://sci-hub.se/https://www.jstor.org/stable/24700283, Lin *et al*. (2013) point out that for very large samples, *P*-levels go quickly to zero and solely relying on *P*-values can lead researchers to claim significance for differences that are of no practical worth.

With this in-mind, transitioning from manually observed thermometers to data-loggers and platinum resistance temperature (PRT) probes, *significance* of the sign-preserved running difference between instruments does not necessarily imply the difference is important. As measurements are subject to error (uncertainty) including that: (i), neither instruments can sample exactly the same parcels of air 100% of the time; (ii), thermometers may be misread;

and (iii), PRT-probes are prone to spiking, a difference between daily observations could be an artefact, or it could be so small as to be immaterial in a day-to-day sense.

Thus, the strength of an argument is not whether a difference is *significant* or *highly significant* but whether it is *important* in the overall scheme of things. There are increasingly-strident calls in scientific literature for the importance (or size) of a difference to be stated as well as a hypothesis test of the significance of an effect.

Importance of differences between instruments is evaluated using *Cohen's d*, which is calculated as the size of the mean difference between instruments (Site/instrument2, minus Site/instrument1, which is the control), divided by the average standard deviation. While units (°C) cancel out, the magnitude of the difference is expressed in standard deviation units and it follows that the higher the *d* value the more important is the difference (see: Fritz at al., 2012, https://sci-hub.se/10.1037/a0024338).

2. Case study Townsville, Queensland (BoM ID 32040 and 32178)

Daily maximum temperature (Tmax) was measured at Townsville airport using thermometers housed in 230-litre Stevenson screens until 8 December 1994 when they were replaced by 60-litre screens (Figure 1). On that day, an automatic weather station (AWS) commenced operating on a 2m-high mound 200m northwest of the meteorological office. Aerial photographs showed that between August 1995 and July 2002 the supposed 'old' site also moved to a position 93m directly west of the office where another 60-litre screen was installed. As the site was still visible in the July 2002 Google Earth Pro satellite image at coordinates provided in comparison-site, site-summary metadata (Latitude -19.2492°, Longitude 146.7647°) evidence that the 'old' site moved is unequivocal.

Site relocations may cause discontinuities in data and relocating the 'old' site and several previous moves including to the western side of the runway in 1969, was not mentioned in Australian Climate Observations Reference Network – Surface Air Temperature (ACORN-SAT) metadata. Instead, ACORN-SAT misleadingly claimed: "*Observations have been made at Townsville Airport since 1942.* ***There are no documented moves until one of 200m northeast on 8 December 1994, at which time an automatic weather station was installed".*** "*Observations at the new site were made under the original number* (032040)*, while the old site continued until December 2000 under the station number 032178*". It is abundantly clear that under the guise of data homogenisation, site moves and changes at Townsville airport were used by Bureau of Meteorology (BoM) scientists, most recently Blair Trewin, to falsely imply that the climate had warmed (See: https://www.bomwatch.com.au/wp-content/uploads/2020/02/Townsville-full-paper.pdf).

Figure 1. Inside the current 60-litre Stevenson screen at Townsville airport. At the front are dry and wet-bulb thermometers, behind are maximum (mercury) and minimum (alcohol) thermometers, held horizontally to minimise "wind-shake" which can re-set the instruments, and at the rear, which faces north, are dry and wet-bub PRT (AWS) sensors. Moistened by a small patch of muslin tied by a cotton wick that dips into the water reservoir, dry-bulb minus wet-bulb T (wet-bulb depression) is used to estimate relative humidity and dew point temperature. (BoM photograph).

So, although inter-site comparison data are available from 9 December 1994 to 31 December 2000 (BoM ID 32178), the manually observed site moved and the screen size changed at the same time as the automatic weather station (AWS) commenced operating with its 60-litre screen at the current site.

2.1 Methods

Data were downloaded from the BoM, aligned manually using Excel, then processed and analysed using R (https://www.r-project.org/). Briefly, daily manual and AWS data were de-seasoned as separate variables by deducting day-of-year (1-366) averages to give daily anomalies, which were differenced (AWS minus thermometer anomalies) as an additional variable. The resulting final dataset consisted of 2,212 complete cases of raw daily data for each location (Site1 (manual) and Site2 (AWS)), de-seasoned anomalies (Anom) for each, and their difference (Delta). Preliminary analysis was undertaken using the statistical application PAST from the University of Oslo: https://www.nhm.uio.no/english/ research/ resources/past/, and should be duplicable using proprietary statistical packages such as Minitab.

Preliminary tabular and graphical analyses using PAST was used to get a feel for the data, with statistical analysis as the subsequent step.

2.2 Results

2.2.1 Preliminary analysis - data properties

The raw data summary (Table 1) shows that differences in means, ranges, measures of variation (standard error, variance and standard variation), quartile distributions etc. between manually observed Site1 data and AWS Site2 data were small.

Table 1. Statistical properties of Site1 (thermometer) and Site2 (AWS) Tmax data used in the study. (Summarised using PAST.)

Property	Site1	Site2
N	2,211	2,211
Min	16.00	15.80
Max	41.60	41.30
Mean	28.83	29.17
Std. error	0.06	0.06
Variance	7.72	7.45
Stand. dev	2.78	2.73
Median	29.10	29.50
25th percentile	27.00	27.30
75th percentile	31.00	31.30
Skewness	-0.37	-0.44
Kurtosis	0.35	0.54
Coeff. Var (%)	9.64	9.36

The mean Tmax effect size is 29.17_{Site2} - 28.63_{Site1} = 0.34; average SD = 2.76, thus *Cohen's d* = 0.34/2.76 = 0.12 standard deviations, which being <0.2 is rated trivial. Graphical analysis showed raw data were strongly cyclic, while in addition to prominent spikes of up to ±6^{o}C, differenced anomaly data exhibited underlying changes and trends caused by factors unknown (Figure 3).

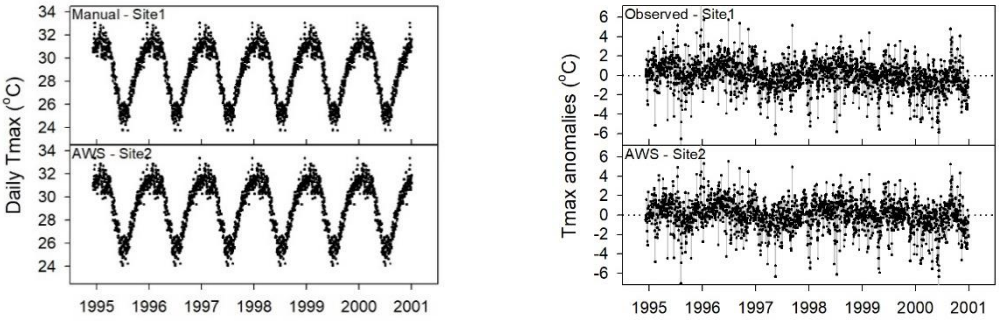

Figure 2. Daily Tmax measured using thermometers in the 60-litre screen, 93m west of the meteorological office (Site1) and Tmax measured by the AWS 200m northwest of the office (Site2). Data are naturally highly variable, and at both sites Tmax anomalies exhibited charges and trends due to unknown factors, and effects due to the weather.

Although instruments were housed in same-sized screens, the sites were 179m apart and potentially affected by impacts and microclimates unique to each site. Data were therefore likely to be confounded with factors unknown. While rainfall may be influential, Figure 3 shows anomaly differences were affected by step-changes, most likely related to undocumented site-changes (including road construction and developments nearby) than changes in the weather, which could reasonably be assumed to be the same across sites.

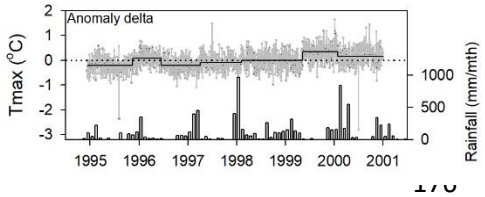

Figure 3. Step-changes in anomaly differences (Site 2 minus Site 1) are indicative of site-change effects. The new mounded site was slightly warmer, particularly after the up-step in May 1999.

2.2.1 Preliminary analysis – seasonality and autocorrelation

Time dependency of one observation on another is determined by the linear correlation coefficient verses the number of periods between times. PAST autocorrelation function (ACF) plots (https://en.wikipedia.org/wiki/Autocorrelation) show repeating seasonal signals in raw data resulted in autocorrelation across all time-lags (Figure 4).

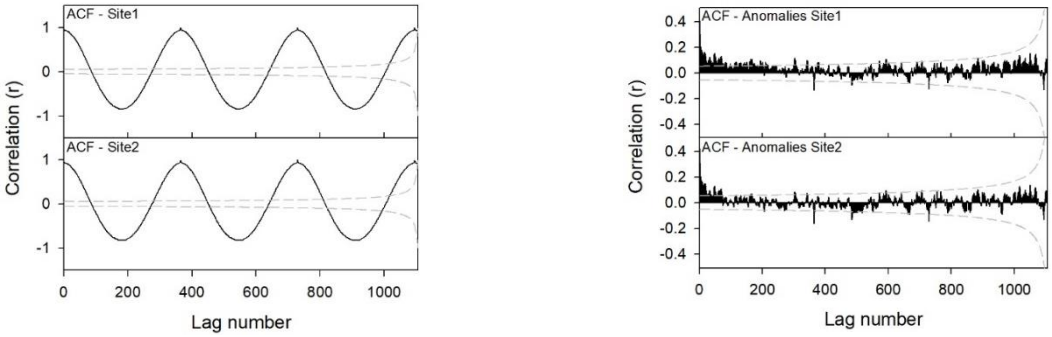

Figure 4. Autocorrelation function (ACF) plot showing correlation between daily Tmax (left) and Tmax anomalies (right) and the same data lagged to the maximum of 1105 days. Grey dashed lines show 95% confidence bands for the linear correlation coefficient (r) within which data are NOT autocorrelated.

By way of explanation, the linear correlation coefficient varies from +1 to -1 (implying negative or positive correlation), with ±1 being a perfect match between data sequences. Grey lines indicate the zone where observed and lagged data would NOT be correlated. Removing the seasonal signal by deducting day-of-year averages from respective day-of-year values considerably reduced autocorrelation between lagged anomalies; however, due to 'hidden'

dependencies, possible trends and site effects (Figure 2) anomalies were still autocorrelated by between 75 and up to 200+ days.

2.2.2 Preliminary analysis – raw data distributions

PAST histogram and normal probability (Q-Q) plots in Figure 5 show: (i), data distributions were not symmetrical (normally distributed) around the Site1 mean of 28.8°C, and that (ii), Site1 data less than about 25°C were cooler than AWS data, while AWS data were warmer above about 31°C (circled).

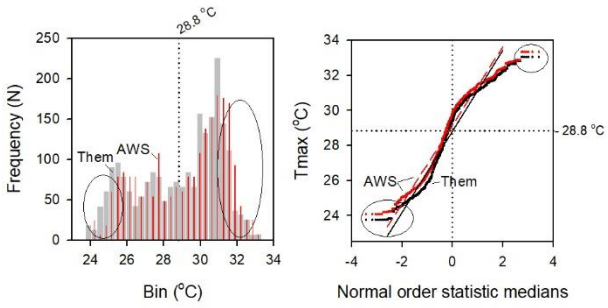

Figure 5. Histograms show Site 1 thermometer data (grey bars) were generally cooler when $Tmax_{Site1}$ was less than about 25°C, while AWS data (Site2, red bars) was warmer where $Tmax_{Site1}$ exceeded about 31°C. Those zones represent the tails of data distributions (circled). The squiggle around the lines in the Q-Q plot on the right results from bimodality (see https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html).

Probability density function (PDF) plots convert frequency histograms, which are stepped, into the *likelihood* of a value occurring within an interval range of one-unit, thereby providing a smoothed representation of the same data. Also, as PDFs are calculated over the same x-axis range and the area under each is unity, the two curves are directly comparable (Figure 6).

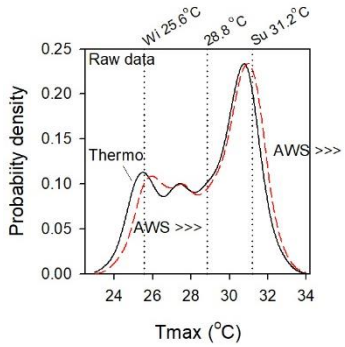

Figure 6. Probability density function plots of the data shown as histograms in Figure 5, confirm that the principal difference between thermometer data at Site1 and AWS data at Site2 occurs in the tails of respective data distributions. Thus, while the mean may be little different, Site2 extremes have shifted warmer relative to Site1.

The bimodal nature of the distributions is due to the sameness of temperatures from June to August (winter), and of higher but similarly static temperatures from December to February (summer).

2.2.3 Preliminary analysis – data distributions of Tmax anomalies

As seasonality, which is a cycle of fixed frequency and amplitude, affects the difference between successive observations advancing and waning across all times, and observations consist of paired data for each day, removing day-of-year cycles from respective day-of-year observations is an essential prerequisite for unbiased analysis. Also, as cycles are predictable their removal should considerably reduce autocorrelation.

Daily anomaly data were more normal in their distribution (Figure 7). However, departure in the Q-Q plot indicates more extreme values in the tails than would be expected if datasets were truly normal. Despite so-called *fat tails*, data were symmetrical, Q-Q plots were parallel, and the normal distribution was a better fit to anomaly data distributions than was the case for raw data.

2.2.4 Preliminary analysis – randomisation and sampling strategies

As an experiment, the dataset was randomised (shuffled) to disrupt dependency of one value on previous values, while the R package *dplyr* (https://cran.r-project.org/web/packages/dplyr/index.html) was used to randomly draw proportions of the

total of 2,211 cases for separate evaluation. *Cohen's d* with 95% confidence intervals was calculated by the *effsize* package (https://cran.r-project.org/web/packages/effsize/effsize.pdf).

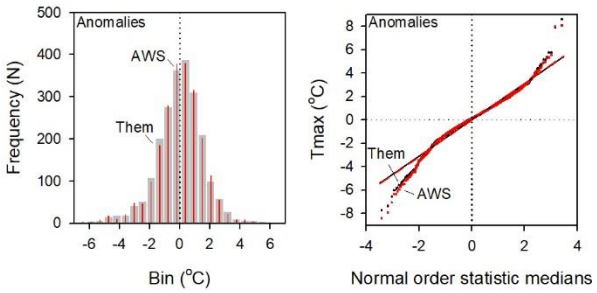

Figure 7. Removing the seasonal signal by deducting day-of-year averages from respective day-of-year data caused datasets to assume a more normal distribution. The 'S'-shaped departure from normality in the Q-Q plot indicated higher number of outliers in the tails of both datasets but otherwise distributions were of similar shape.

2.3 Statistical outcomes

Statistical outcomes are summarised in Table 2.

Paired and un-paired t-tests detected significant differences between sites/instruments in both time-ordered and shuffled raw data, with Site2 being warmer on average by 0.3 to 0.4°C. *P*-levels were also considerably smaller (i.e., more significant) for paired t-tests than unpaired tests, for instance, for N=24 re-ordered random samples $P_{unpaired}$ = 0.466 (not significant) while for the same sample of data-pairs P_{paired} = 2.76e-05 (highly significant).

Despite problems caused by autocorrelation and non-normality, *Cohen's d* calculated that the AWS at Site2 was 0.12 to 0.13 standard deviations warmer than Site1, with the difference ranked as negligible. This reinforces that effects detected as significant due to large sample sizes should not be overvalued as being meaningful or consequential.

Table 2. Paired and un-paired t-tests for raw data and day-of-year anomalies. As differences between randomly sampled paired or un-paired anomaly datasets were not significant, results for those tests are not given.

Comparison Site1 vs Site2	Delta (Site2 - Site1) (°C)	Significance (*P*)	Importance (Cohen's *d* (95% Ci))	Size effect[4] (Magnitude)
Raw data paired and un-paired[1]	0.335	<0.001	0.13 (-0.192, 0.074)	Negligible
Raw data paired and un-paired, shuffle[2]	0.335	<0.001	0.13 (-0.192, 0.074)	Negligible
Anomalies paired and un-paired	5.6-E17	ns (*P* >0.05)	NA	NA
Raw data subsample 2[3]	0.350	0.014	0.13 (-0.231, -0.025)	Negligible
Raw data subsample 3	0.338	0.018	0.12 (-0.227, -0.021)	Negligible
Raw data subsample 1	0.332	0.028	0.12 (-0.218, -0.013)	Negligible

Notes:

[1] For all comparisons, significances were higher for paired verses un-paired t-tests

[2] While shuffling removed autocorrelation it made no difference to test outcomes. It should be noted therefore that autocorrelation in input data affects validity of the test, not its significance.

[3] While data were randomly subsampled, sample size in all cases was N=729

[4] The size effect is assessed using the thresholds provided in (Cohen 1992, updated in 1988), *viz*. |d|<0.2 "negligible", |d|<0.5 "small", |d|<0.8 "medium", otherwise "large".

Citation: Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York:Academic Press.

2.3.1 The effect of sample size on the significance of unpaired t-tests

Akin to a Monte Carlo simulation, site differences and *Cohen's d* was evaluated by randomly sampling progressively larger numbers of cases (with replacement) from an initial 1%/year (N=24), advancing by 2%/year, to N 1740 in 40-rounds, which represented 78% of the dataset (Figure 8). Samples were not time-ordered and a duplicate experiment showed the same result. (If data were re-ordered, autocorrelation emerged after 3-rounds i.e., when the number of time-ordered samples equalled or exceeded approximately N=109.)

While the t-statistic for the unpaired t-test declined (became more significant) from *P*=0.69 at N=24, to *P*=0.05 at N=560, after steading at N=332 response variables showed no marked change within the bounds of sampling variation. Thus, it could not be claimed that increasing *significance* (i.e., declining *P*-levels) were related to increased differences between dataset means or changed effect sizes.

Dependence of *P*-level on N and not on the response variable potentially results in Type1 error, which is declaring differences to be significant when they may not be; or that the significances could be a test artefact. For instance, the same 0.33°C to 0.34°C difference which was not significant below N=560, became significant because the pooled standard error declined as sample size increased.

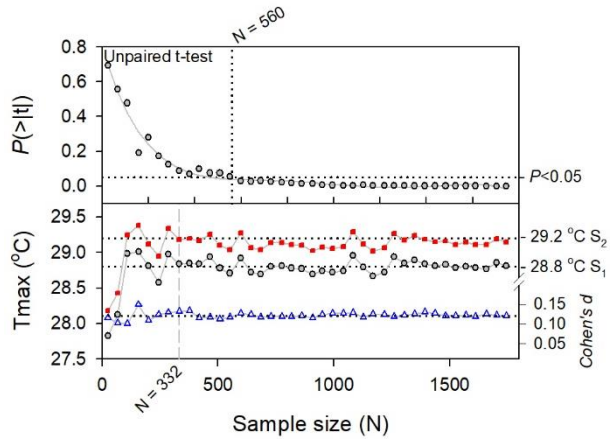

Figure 8. As sample size increased from N=24 to N=560, *P* declined (became more significant), while after steadying at N=332, differences between Site1 and Site2 remained the same. Averaging 0.12 standard deviations, *Cohens d* (effect size, blue triangles) was also not responsive to sample size.

2.4 Discussion

This note outlines in general terms, assumptions underlying the use of paired and unpaired t-tests for comparing timeseries of maximum temperature data measured in parallel by different instruments housed in the same or different Stevenson screens. The overlap dataset used as the case study consisted of same-day manually observed thermometers and AWS-probes, housed in 60-litre screens located 179m apart at Townsville airport from 9 December 1994 to 31 December 2000. The research question is whether Tmax observed by the different instruments at the two sites was different.

A vexing question at the outset, is whether data collected each day by separate instruments in same-sized screens or co-located in the same screen (e.g., Figure 1) represent two *independent* subject groups (observations with no connection between them), or if data represent true data-pairs collected sequentially from homogeneous subjects, items or things as required for a valid repeated-measures (paired) t-test. To draw an analogy, a paired t-test would be valid if two instruments were used on each of a succession of subjects; however, if each instrument measured subjects that were not precisely the same and subjects and instruments formed separate groups, an un-paired t-test would be appropriate and the paired t-test would not.

While seemingly pedantic, the question is important because as pairing is intended to control variation *within subjects*, significance levels (*P*) are considerably enhanced relative to comparing means of the same data using an un-paired test.

The two instruments either in two separate screens or co-located as in Figure 1, with one nearer the rear of the screen and the other near the front, are the test subjects, while daily Tmax is the response variable which the paired design assumes to be homogeneous (exactly the same for each instrument each day). However, irrespective of whether data are available each day, air within or between screens is unlikely to be spatially and timewise homogeneous,

thus testing differences between independently observed populations of values would be more appropriate than paired t-tests under the circumstances.

Furthermore, a Monte Carlo sampling experiment of pairs of raw data found the paired test needed less than 24 randomly chosen cases to find significance, while the unpaired test needed 560 data-pairs. Bias resulting from choosing one test that may not be appropriate, over another may therefore be considerable.

As annual day-of-year cycles dominated both datasets and raw Tmax was autocorrelated at all lags, comparing raw data using paired or unpaired t-tests was invalid anyway. Autocorrelation arises because average Tmax cools toward the end of each month during the cooling phase from February to June, and the end of each month gets warmer during the warming phase, which explains reversal (from positive to negative correlation) of the ACF plot (Figure 3).

Autocorrelation is a property of the data not the test. It causes estimated standard errors (the noise, which is the dominator) to be underestimated relative to differences (the signal, which is the numerator), which increases the *t*-value and thus the likelihood of significant effects. Testing for autocorrelation and removing underlying seasonal cycles by deducting day-of-year averages from respective day-of-year data maintained timewise integrity of the data, including relationships with covariables. It also removed bimodality, caused datasets to assume a more normal distribution and vastly reduced the magnitude of uncontrolled factors including lagged processes (unaccounted-for variables), without affecting data-properties. Whether compared using paired or un-paired t-tests, removing seasonal cycles also caused Site2 data to be not significantly different to data for Site1. While Figure 6 shows tails of Site2 data distributions were warmer, that the raw data means were different was an artefact of the tests.

Finding statistical outcomes were the same for time-ordered as they were when autocorrelation was disrupted by shuffling (Table 2) pointed to another issue, which is the effect of the large sample size (2,212 cases) on finding diminishingly-small differences as highly significant, when they were not meaningful.

The random sampling exercise (Figure 8) showed the weaknesses of analysing increasingly large sample sizes using un-paired t-tests. It was found that as the number of samples increased, pooled standard errors became diminishingly small, so the significance of the test increased independently of the difference or its importance (effect size). Furthermore, applying the wrong test (the paired t-test) to randomly selected data-pairs vasty increased the likelihood of detecting spurious differences. Data-shopping in all its forms, including using inappropriate tests undermines trust in the outcome.

A suggested protocol for undertaking preliminary investigations of paired datasets is given in Appendix 1 and the accompanying Excel workbook contains a worked example of calculating and deducting day-of-year averages from respective day of year data.

Conclusions

It is vital at the outset of undertaking paired and un-paired t-tests of intensively sampled timeseries to examine properties of datasets and mitigate the presence of autocorrelation. As the significance of test outcomes also increase as the numbers of samples increase, solely relying on *P*-values can lead researchers to claim significance for differences that are of no practical worth. Researchers are therefore encouraged to employ an empirical measure of whether a *significant* or *highly significant* outcome is meaningful.

Notwithstanding the problem of misleading ACORN-SAT metadata, which must be deliberate, with effects of autocorrelation and non-normality much diminished, neither paired nor

unpaired t-tests detected significance in the difference between Tmax measured at Site1 using thermometers and Site2 by the AWS at Townsville airport.

Dr. Bill Johnston

3 June 2023

Disclaimer:

This note is intended to provide guidance of a general nature specific to undertaking comparisons between meteorological instruments. While the Author undertook an undergraduate course in biometry, and post-graduate workshops etc., and has since honed those skills through reading, investigation and practical application using R, he does not claim to be a statistician.

Acknowledgements

Editorial assistance provided by David Mason-Jones is greatly appreciated.

Appendix 1

A suggested protocol for undertaking investigations of paired datasets using PAST from the University of Oslo: https://www.nhm.uio.no/english/ research/ resources/past/

(Citation: Hammer, Ø., Harper, D.A.T., Ryan, P.D. 2001. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* 4(1): 9pp.)

Paste data into PAST *[PAST manual* (v. 4.08) *p. 10]*

Firstly, calculate summary statistics and check the treatment (instrument) means and moments (1st Quartile (25th pc), median and 75th Q); also, standard error, standard deviation (SD) and variance (which are related) *[PAST manual p. 47].*

> Calculate the difference between the means as a ratio of the standard deviation using the formula: (SiteB-SiteA)/SD, this gives *Cohen's d* which robustly determines if the difference (the effect size) is likely to be negligible (less than 0.2 SD units), small (>0.2), medium (>0.5) or large (<0.8).

Secondly, plot data in time order (as graph, not *x,y* graph) and examine cycles and trends. *[PAST manual p. 24]*

Thirdly, make a histogram (graph), overlay the normal distribution line and look to see if data are normally distributed, bimodal, long or short tailed etc. *[PAST manual p. 27. Note, change graph properties to emphasise aspects of data; numbers can be copied.]*

Fourth, plot a Q-Q (normal distribution) plot and decide if data need to be adjusted or transformed. *[PAST manual p. 33; numbers can be copied.]*

Fifth, plot an ACF (autocorrelation) plot to identify signals in the data and how they are related (i.e., check data are independent, or if the degree of AC is of concern. *[PAST manual p. 211; numbers can be copied.]*

Calculate and remove day-of-year cycles from each timeseries. (See method and formulae provided in the Townsville data workbook (Townsville_PAST.xlsx).

1. Prepare the data as per DataPrep tag, including pairing, calculating dates and removing rows having missing observations. Headers are in Row3, data start in Row4)
2. Insert index column to the left of the data (Col A, DayNum) that indexes day of the year (DayNum) for all pairs of observations (1 to ~ 366). This can be done in Excel using the date column (Col B) as: A4=B4-DATE(YEAR(B4),1,0); copy to the bottom of the datatable.
3. Make a pivot table based on Day of Year, and set the values to be averages for each timeseries. Copy the pivot table and paste *as numbers* a few columns to the right of the datatable, say in Columns J (Row label), K (Site1), and L (Site2).
   [The three-column list forms a lookup table showing the average for each of day of year (366-rows of data).] Insert a Column after Date (Col C) and calculate DeciYear (for graphing). The formula is C4 = YEAR(B4)+A4/365.25
4. Assuming data are in the order of: DayNum, Date, DeciYear, Site1 and Site2 (Cols A to E) with the first data row being Row4; with the lookup table in Columns J [day number], K [Site1 day-of-year average] and L [Site2 day-of-year average], the lookup formula is:
   =C4-VLOOKUP($A4, $J$4:$L$369,2) [Paste formula into F4; Name Col F as Site1Anom]
   =G4-VLOOKUP($A4, $J$4:$L$369,3) [Paste formula into G4; Name Col F as Site2Anom]

[The command looks up the reference row number in Col A, in the first column of the lookuptable array [Col J] and return the value in the second column of the array [which is Col K] specified as ,2) and also the value for ,3), for the third array column.]

Respective formulae will then deduct the lookup value, from the data values in Col D and E to give daily anomaly values in Cols F and G. Copy the formulae to the bottom of the datatable, name the columns *Site1Anom* and *Site2Anom* and calculate Delta (Site2Anom minus Site1Anom). It is advisable to open a new worksheet and copy and paste the DataTable **as numbers** into the new sheet, then save the Workbook.

Paste the anomaly data into PAST and repeat the first four steps using anomaly data i.e., summary, timewise graph, histogram, Q-Q plot and ACF plot and check how things changed.

You can either then continue with PAST and do Univariate samples, two-sample tests *[PAST Manual p. 53]* and two-sample paired tests *[PAST Manual p. 62]*, or with Minitab, or whatever. If you have the option (which you don't in PAST (yet)) ask for Cohen's d with 95% confidence intervals.

As anomalies have been de-cycled, it is important to compare if tests on raw data are still significant when conducted using anomalies.

3 June 2023