Analysis of parallel temperature data using t-tests

Part 2. Brisbane airport

Dr Bill Johnston

www.bomwatch.com.au

Summary

At the outset of undertaking comparisons between instruments housed in the same or different Stevenson screens or between different sites, a robust protocol that evaluates properties of datasets should be used to guide and validate statistical outcomes. While small differences that are not meaningful become increasingly significant as the number of samples increase, the effect size statistic (*Cohens d*) provides an initial indication whether a difference between dataset means (ΔTmax/standard deviation) is likely to be important. *Cohens d* less than 0.2 indicates the effect size is negligible in the overall scheme of things.

T-tests are not suitable for comparing differences in the means of closely spaced time-series unless steps are taken to mitigate autocorrelation which results from embedded cycles and possible trends related to weather or other uncontrolled factors. Cycles may be removed from daily data by subtracting day-of-year (1-366) averages from respective day-of-year data to derive de-seasoned anomalies. These may be differenced to highlight discontinuities caused by a third factor such as changed instrument responses.

The paired t-test explicitly presumes all variation in the response is attributable to subjects. However, meteorological instruments held within Stevenson screens are unable to sample the same parcels of air 100% of the time, consequently some or all of the variation between instruments is attributable to the air being measured. Even if the use of t-tests is valid (for example by random sampling), the Type1 error-rate (declaring a difference to be significant when it is not) for the same data is considerably higher for paired t-tests than the un-paired test.

1. Background

Use of paired t-tests to analyse time series of daily weather observations, discussed recently at https://wattsupwiththat.com/2023/05/26/bureau-capitulates-but-overseas-model-unlikely-to-solve-all-temperature-measurement-issues/, highlights the need for a protocol-based approach to comparing daily maximum temperature (Tmax) observed using thermometers with data from automatic weather stations (AWS). While Tmax thermometers record a single observation each day, AWS derive Tmax from a stream of high frequency platinum resistance temperature (PRT) probe samples. Depending on the purpose, comparisons may involve different instruments at two different sites, or two instruments co-located in the same Stevenson screen, or separately side by side at one site. Instruments and sites are compared to determine if a change was likely to affect the trajectory of on-going observations and to derive adjustment coefficients.

Using parallel data for Townsville airport from 9 December 1994 to 31 December 2000 a protocol was advanced to ensure outcomes of tests are valid. In particular that autocorrelation, which is the interdependence of values at one time on observations for previous times is mitigated. Also, as significance increases exponentially as the number of daily data pairs

increase, an empirical measure of whether differences are meaningful assists interpreting statistical outcomes. Biases arising from using paired verses unpaired t-tests for comparing instruments that cannot measure the same parcels of air 100% of the time was highlighted.

2. Brisbane airport overlap

Brisbane airport is an Australian Climate Observations Reference Network – Surface Air Temperature (ACORN-SAT) site used to monitor Australia's warming. It had previously been homogenised by Simon Torok to include Brisbane Regional Office (ID 40214) - a site that apparently operated in the Brisbane Botanical Gardens, but the RO-site was not used by ACORN-SAT. (The true location of the Brisbane Regional Office site appears to be unresolved.)

According to the ACORN-SAT catalogue the original airport meteorological site moved from the vicinity of the Eagle Farm railway station (closed in 1993) about 800m north in August 1955 to a position now occupied by the Gateway Motorway. An AWS installed at the southern end of the new airport main runway on 30 November 1987, 3 km east of the previous site became the primary instrument on 1 November 1996. The current site (040842, Figure 1) was established earlier in 1994, 3.3 km to the northeast of the previous one (040223), which closed in 2000.

During the period between when the current site opened (1 April 1994) and the previous site closed (6 February 2000), both sites operated in parallel using AWS and 60-litre Stevenson screens. The overlap dataset is used in this study to evaluate utility of protocols suggested previously for the site/instrument comparison for Townsville and broadly follows that format.

Figure 1. A view of the current site at Brisbane airport, photographed by the BoM on 9 February 2012 (from the ACORN-SAT Catalogue). The Stevenson screen housing instruments beyond and slightly to the right of the Dynes pluviometer in the immediate foreground. Now superseded by a tipping-bucker raingauge, the copper-clad Dynes was a chart-recorder that measured rainfall intensity.

2.1 Methods

Data for the two Brisbane airport sites were downloaded from the Bureau of Meteorology climate data online facility, aligned manually using Excel and processed and analysed using R (https://www.r-project.org/). Briefly, datasets were de-seasoned as separate variables by deducting day-of-year (1-366) averages to give daily anomalies. Anomalies were differenced (current minus previous, which was the control) as an additional variable. The resulting final dataset comprised 2,138 incomplete cases of raw daily data for each of two sites (Site1 (previous) and Site2 (current)), de-seasoned anomalies (Anom), 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 including Minitab.

2.2 Results

Preliminary tabular and graphical analyses was used to get a feel for the data, and to guide subsequent statistical analysis.

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 the previous AWS (ID 40223) and current site (40842) were small. The approximate effect size of the difference (*Cohens d*) in raw data means ($Delta/SD_{Av}$) is rated negligible (i.e., <0.2). As anomalies are zero-centred *Cohens d* is unavailable.

Table 1. Statistical properties of Tmax data used in the study (summarised by PAST.) Site1 refers to the comparator (ID 40223) which closed in 2000, Site2 to the current site (40842), which is on-going. During the comparison period both sites operated AWS and 60-litre Stevenson screens.

	Previous site	Current site	Previous site	Current site	
Statistic	Site1	Site2	Site1Anom	Site2Anom	DeltaAnom
N	1,918	2,138	1,,918	2138	1,918
Min	14.10	14.30	-7.28	-7.30	-2.05
Max	37.90	39.10	10.42	8.75	3.17
Mean	24.84	25.11	0.00	0.00	0.02
Std. error	0.08	0.08	0.04	0.04	0.01
Variance	12.84	12.78	2.99	3.43	0.37
S.D.	3.58	3.57	1.73	1.85	0.61
Median	24.80	25.00	-0.03	-0.07	-0.03
25 prcntil	22.10	22.40	-1.03	-1.13	-0.36
75 prcntil	27.70	27.80	1.00	1.05	0.35
Skewness	0.08	0.09	0.39	0.37	0.60
Kurtosis	-0.50	-0.33	2.25	1.44	1.89
Coeff. var	14.42	14.24	na	na	na
Effect Size	$0.27_{Delta}/3.58_{SD}$	d = 0.07	na	na	na

Graphical analysis showed raw data were strongly cyclic, while in addition to prominent spikes of up to ±6°C, anomaly data exhibited underlying changes and trends due to the weather and other possible factors (Figure 2).

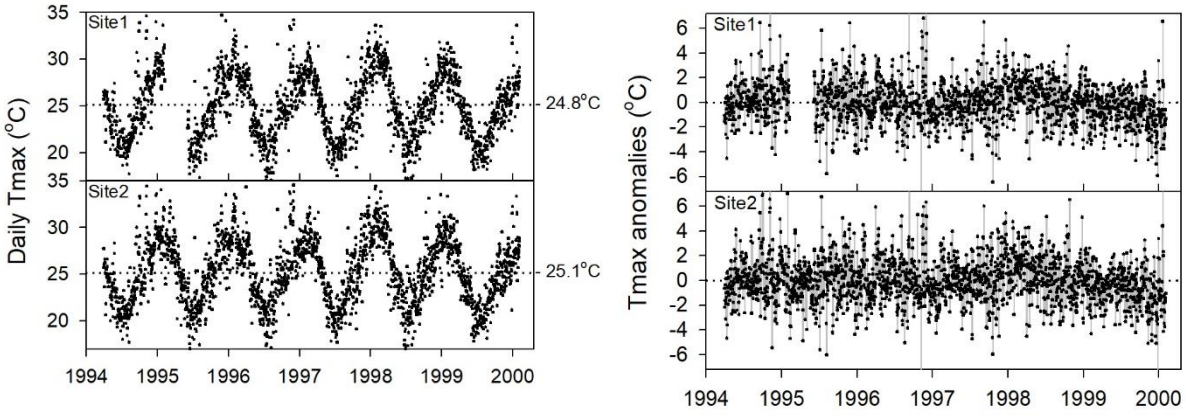

Figure 2. Daily Tmax at Site1, and at the new site 3.2km north at Site2. Data are naturally highly variable and at both sites, Tmax anomalies (right) exhibited charges and trends due to due to the weather and possibly unknown factors unique to each site.

2.2.2 Preliminary analysis – seasonality and autocorrelation

Time dependency of one observation on another is determined by the linear correlation coefficient for the numbers of periods (lags) between times. PAST autocorrelation function

(ACF) plots (https://en.wikipedia.org/wiki/Autocorrelation) show the repeating seasonal cycle resulted in autocorrelation across all time-lags at both sites (Figure 3). Although autocorrelation was considerably reduced by removing the dominant seasonal signal, 'hidden' dependencies, possible trends and site effects mentioned previously in relation to Figure 2 resulted in autocorrelation up to lags of 100 to 150 days.

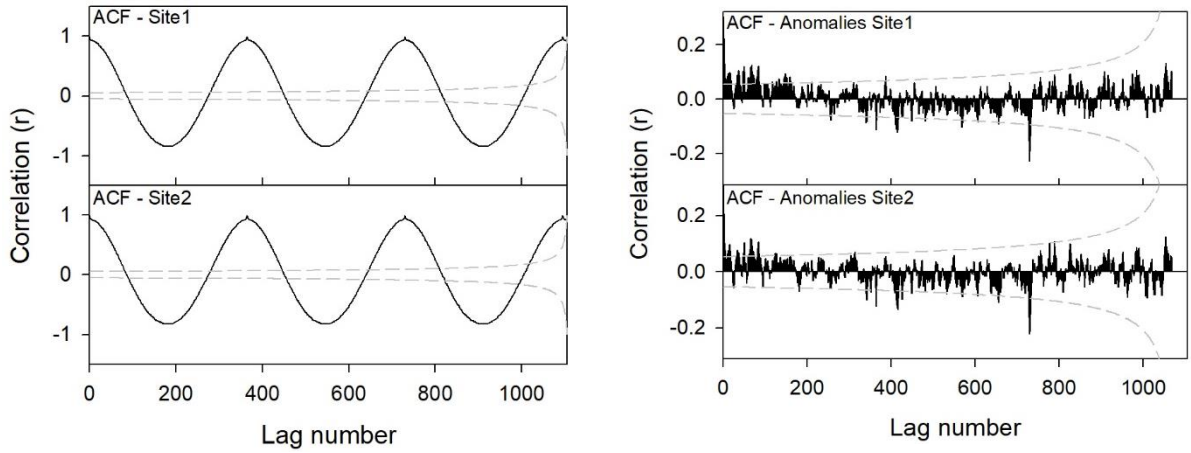

Figure 3. Autocorrelation function (ACF) plot showing correlation between daily Tmax (left) and Tmax anomalies (right) and the same data lagged to the maximum of 1105 days. Grey dashed lines show 95% confidence bands for the linear correlation coefficient (r) within which data are NOT autocorrelated.

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 (i.e., the r value is less than the critical value for that lag, thus $P_{uncorrelated}$ >0.05).

Although both sites were monitored by AWS operating with same-sized screens, they were 3.2km apart and potentially affected by impacts and microclimates unique to each site. Data were therefore likely to be confounded with unknown factors. While rainfall may be influential, particularly the 1997-1998 El Niño, Figure 4 shows anomaly differences were affected by step-changes, the most obvious of which from 1 February to 6 June 1995 was most likely related to an undocumented site-change at Site1.

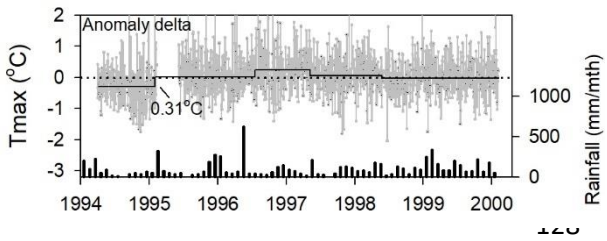

Figure 4. Step-changes in the mean of Site2 minus Site1 anomalies. The break in Site1 data from 9 February 1995 to 6 June 1995 and the subsequent up-step suggests the site changed at that time, but there is no mention of a disturbance in site-summary metadata. Remaining changes could be weather related.

2.2.3 Preliminary analysis – raw data distributions

PAST histogram and normal probability (Q-Q) plots in Figure 5 show data distributions were not symmetrical (normally distributed) around the Site1 mean of 24.8°C. Site1 was cooler than Site2 at Tmax less than about 20°C, and Site2 was warmer above about 27°C (circled) and in the warm-tail of the distribution.

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 resulting in continuous distribution curves. 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).

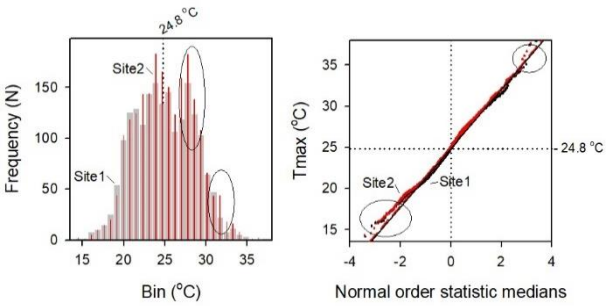

Figure 5. Site 1 data (grey bars) were generally cooler when $Tmax_{Site1}$ was less than about 20°C. Site2 was warmer where $Tmax_{Site1}$ exceeded about 27°C (circled). Those zones represent the tails of data distributions. Except for the tails, the Q-Q plot on the right indicates data are acceptably normal (see https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html).

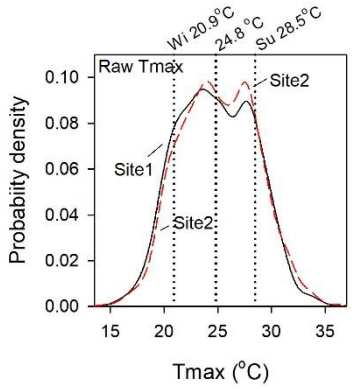

Figure 6. Probability density function plots of the data shown as histograms in Figure 5, confirm that the principal difference between Site1 and Site2 occurs in the tails of respective data distributions. Thus, while the mean may be little different, Site2 extremes appear to have shifted slightly warmer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

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

2.2.4 Preliminary analysis – distributions of Tmax anomalies

Seasonality, which is a cycle of fixed frequency and amplitude, affects the difference between successive observations as cycles advance and wane through time. Removing the cycle by deducting day-of-year averages from respective day-of-year observations is an essential prerequisite for unbiased analysis. Further, as they 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 data were skewed right (see: https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html). Despite so-called *fat tails*, data were symmetrical, Q-Q plots were parallel, bimodality was removed and the normal distribution was a better fit to anomaly data distributions than was the case for raw data.

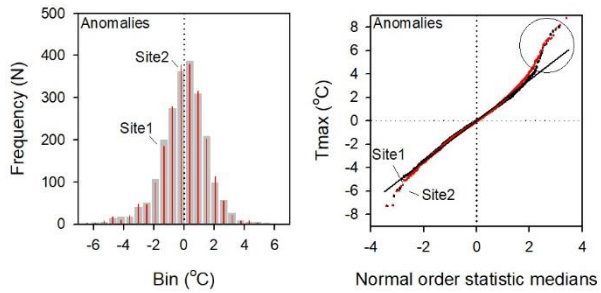

Figure 7. Removing the seasonal signal by deducting day-of-year averages from respective day-of-year data caused datasets to assume a more normal distribution. The 'S'-shaped departure from normality in the Q-Q plot indicated data were skewed right (i.e., although of similar shape, data for both sites were warm-skewed: circled).

A PDF plot of data shown as a histogram in Figure 7, confirms that while the probability of daily values is higher at Site1 near the centroid of the distribution (where x = 0), upper-range-Tmax was slightly warmer at Site2 (Figure 7b).

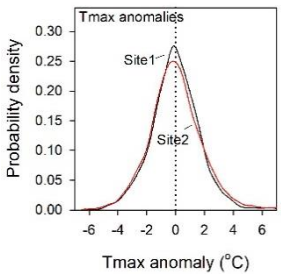

Figure 7b. Probability density plot of the histogram shown in Figure 7.

While Figures 7 and 7b appear to show site-related differences, they may be too slight to be significant. Tests available in PAST including the Mann-Whitney and Mood tests for equal medians, and the Kolmogorov-Smirnov, Anderson-Darling and Epps-Singleton tests for equal distributions found no significant differences between sites. Differences therefore reflect random rather than systematic behaviour.

2.2.5 Preliminary analysis – randomisation and sampling strategies

As an experiment, the dataset was randomised (shuffled) to disrupt dependency of one value on previous values. The R package *dplyr* was used to randomly draw proportions of the total of 2,211 complete cases for separate evaluation (https://cran.r-project.org/web/packages/dplyr/index.html). *Cohen's d* with 95% confidence intervals was calculated by the *effsize* package (see https://cran.r-project.org/web/packages/effsize/effsize.pdf).

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 0.24°C. *P*-levels were also considerably smaller (i.e., more significant) for paired t-tests than unpaired tests. For instance, for 56 randomly drawn samples $P_{unpaired}$ = 0.712 (not significant) while P_{paired} = 0.0006 (highly significant). The paired test significance level approached the machine limit at N= 250 (*P* = 7.795e-10 i.e., 10 decimals before the numeral), while for the same samples, $P_{unpaired}$ = 0.416.

Calculated *a priori*, *Cohen's d* indicated that the AWS at Site2 was 0.07 to 0.06 standard deviations warmer than Site1, with the difference ranked as negligible (Table 1). Effects detected as significant due to large sample sizes should therefore not be overvalued as being meaningful or consequential in the overall scheme of things.

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

Site differences and *Cohen's d* was evaluated by randomly sampling progressively larger numbers of cases (with replacement) from an initial 1%/year (N=17), advancing by 2%/year, to N=1901 in 50-rounds, representing 82% of the dataset (Figure 8). Samples were not time-ordered and therefore were not autocorrelated and a duplicate experiment gave identical results. If data were re-ordered, autocorrelation emerged after 2-rounds when the number of samples equalled or exceeded about 56.

Figure 8 illustrates the statistical fallacy that as the numbers of samples increase, small differences that are not meaningful, and which could be due to accumulated outlier values or averaging beyond the precision of the dataset, become increasingly significant. While the difference between sites stabilised at about N=500, it was not until N=1600 that it became significant at *P*=0.5. Thus, it could not be claimed that increasing *significance* (i.e., declining *P*-levels) were related to increasing differences between datasets or changed effect sizes as N increased.

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[1]	0.243	<0.001	0.07 (0.135, 0.008)	Negligible
Raw data un-paired	Ditto	0.037	Ditto	Ditto
Raw data paired shuffle[2]	Ditto	<0.001	Ditto	Ditto
Raw data un-paired, shuffle[2]	Ditto	0.037	Ditto	Ditto
Anomalies paired and un-paired	5.6-e17	ns (*P* >0.05)	NA	NA
Raw data subsample 1[3]	0.258	0.206 (ns)	0.07 (0.181, 0.039)	Negligible
Raw data subsample 2	0.204	0.772 (ns)	0.06 (0.166, 0.054)	Negligible
Raw data subsample 3	0.258	0.186 (ns)	0.07 (0.184, 0.036)	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, which depended on the sample size. 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=634

[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.

Dependence of *P*-level on N and not on the response variable potentially results in Type1 error, which is declaring a difference to be significant when it is not. In this case, the same 0.21°C to 0.26°C difference which was not significant below N=560, became significant because the dominator in the test equation (the pooled standard error) declined as sample size increased. Consequently, the size of the t-statistic increased, the *P*-level declined and significance was attained at N=1600 for an immaterial difference of 0.24°C.

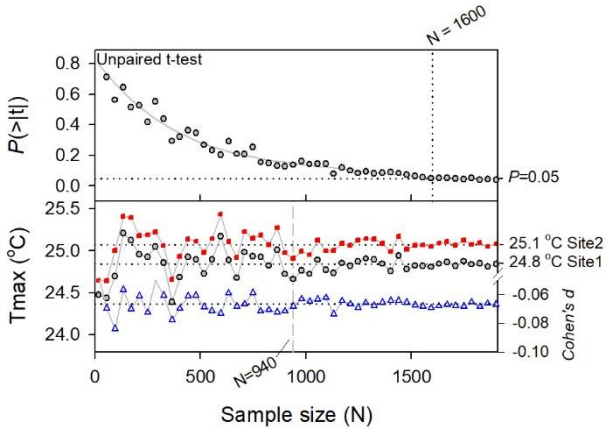

Figure 8. The effect of sample size on significance levels (*P*>(|t|)), mean-Tmax for Site1 and Site2 (grey circles and red squares), and *Cohens d* (triangles, right axis). Significance levels increased as *P* declined reaching *P*=0.05 at N=1600 data pairs. While sampling variation resulted in noisy data up to N=940, the difference of 0.24°C was mostly unchanged. (Note that the site means were rounded to 1-place). ACF-plots showed the t-test of randomly drawn data pairs was not affected by autocorrelation.

2.4 Discussion

This study used the overlap dataset for Brisbane airport from when the current site opened on 1 April 1994 and the previous site closed on 6 February 2000, to evaluate the strengths and weaknesses of protocols advanced previously using the site/instrument comparison for Townsville.

While data for each site comprised one observation per day, it was unlikely that AWS PRT-probes housed in 60-litre Stevenson screens 3.2km apart could sample the same parcels of air 100% of the time. Thus, this study broadens the previous investigation of Townsville data to a situation where site differences were more likely.

Paired t-tests assume that differences in responses are strictly attributable to subjects as would be the case where instruments were compared under controlled conditions, simultaneously such as in the same oil- or ice-bath. However, as air circulates through a Stevenson screen randomly and without spatial control, conditions during the heat of the day are turbulent and changeable. Even if housed in the same screen, two instruments are unlikely to sample precisely the same parcels of air 100% of the time.

The t-test measures the ratio of signal to noise, in this case the difference between instruments divided by variation as measured by the standard error (pooled in the case of unpaired tests) (see for example: https://www.ugr.es/~fmocan/MATERIALES%20DOCTORADO/Statistics-for-Clinicians-4-Basic-Concepts-of-Statistical-Reasoning-Hypothesis-Tests-and-the-T-test.pdf). As the denominator in the test equation is less for paired-tests than if the same data were analysed using the un-paired or two-sample test, significance differences are more likely. Data measured on the same day does not mean they are paired in the sense required by the test and using a test inappropriately on the basis that significance levels are higher is misleading.

A key question is how many samples (observations) are needed to detect a Tmax difference between Site1 and Site2 at Brisbane airport? Also, what is the effect of analysing data using paired t-tests (i.e., as paired differences) versus the more appropriate two-sample unpaired test (Site1 vs Site2).

Table 3 shows the Tmax difference as *Cohens d*, which is expressed in standard deviation units. As mentioned previously, regardless of whether the difference is significant, values <0.2 are trivial or meaningless in the overall scheme of things. A meaningful difference would fall between 0.5 and 0.8 SD units (footnote to Table 2). So, what would that look like, holding Site1 Tmax at its mean, compared to a hypothetical dataset having similar statistical properties (number of samples and standard deviation) but progressively increasing temperatures?

Table 3. Average Tmax for Site1 and Site2, 95% bootstrapped confidence intervals, differences between means (Delta) as a ratio of the pooled standard deviation (*Cohens d*). Note that the upper 95% CI for site1 (25.0°C) overlaps the lower 95% CI for site2 (24.96°C).

Statistic	Site1 (°C)	CI_{95} Lo	CI_{95} Hi	Site2 (°C)	CI_{95} Lo (°C)	CI_{95} Hi (°C)	Delta (°C) (mean[1])	*Cohens d*
Mean	24.84	24.68	25.00	25.11	24.96	25.26	0.27	
Stdev	3.58			3.57				
Pooled	3.58							0.075

[1] Delta is calculated from raw data, not complete pairs as was the case for t-tests

A Monte Carlo experiment was conducted using R, whereby N=1918 normally distributed random numbers, having a mean of 24.84°C and standard deviation (Stdev) of 3.58 (Table 3) were paired with a second series (Series2) having the same Stdev and number of samples but with Tmax increasing incrementally. Tmax for Series2 increased from 25.0°C (the upper 95% bootstrapped confidence interval shown in Table 3) in 0.15°C increments until *Cohens d* equalled 0.8. Although numbers of samples were excessive, data were random and not autocorrelated.

At each iteration, paired and unpaired t-tests were conducted and *Cohens d* was calculated using the *effsize* package (https://cran.r-project.org/web/packages/effsize/effsize.pdf). Results are shown in Figure 9.

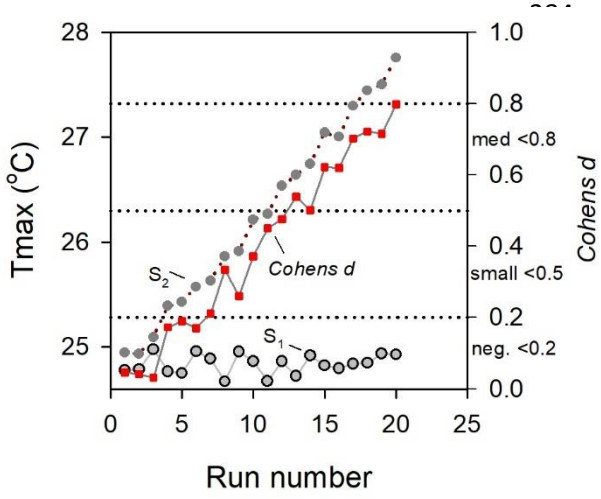

Figure 9. Twenty simulated rounds of normally distributed Series1 (S1) data with mean=24.84°C, Sdev=3.58 and N= 1918 (Table 3), and a second series (S2), where with Sdev and N held the same, mean Tmax increased from 25.0°C (the upper 95% bootstrapped confidence interval in Table 3) until *Cohens d* (right axis) equalled 0.8.

*W*hile the simulated difference between Series1 and Series2 was significant after 3-rounds (*P*=006), the effect size was negligible until *Cohens d* approached 0.5 and the difference exceeded 1.6°C.

Probability density plots for several of the same runs contributing to Figure 9, are presented in Figure 10. As all plots were calculated over the same x-axis range, they are directly comparable.

Taken together, Figures 9 and 10 show that it is not until the effect size exceeded 0.2 (small effect, 6-rounds), that simulated distributions showed a clear separation such that the difference between Series1 and Series2 of 0.62°C could be regarded as both *significant* and *meaningful* in the overall scheme of things.

Likewise, a power analysis using the R package *pwr* (https://cran.r-project.org/web/packages/pwr/pwr.pdf) found that for the raw data effect size of 0.075, the optimum sample size for a two-sided paired t-test was N=1,8175 daily data pairs, while for a two-sample test, N=36,347!

From multiple perspectives (numbers of samples and effect size (Figure 8), bootstrapped confidence intervals (Table 3), and Monte Carlo scenario comparisons (Figures 9 and 10)) the difference between sites is too small to be meaningful. Nevertheless, provided test-assumptions are ignored, significance can be achieved if sufficiently large numbers of samples are available. The trade-off between significance and effect size is central to avoiding the trap of drawing conclusions based on statistical tests alone.

Critical to the issue of comparing instruments and sites is that the correct statistical test is used and that assumptions are not violated. Of particular importance is that paired differences, in the case of paired t-tests, and for un-paired tests, that data within groups are not autocorrelated by embedded cycles and trends (Section 2.2.2). If having removed cyclic components by deducting day-of-year averages autocorrelation remains, tests require subsets to be drawn randomly, or heteroskedasticity and autocorrelation-consistent (HAC) standard errors to determine if differences are significant.

The second major issue is that small differences that are not meaningful become significant as the number of samples increase. Random sampling from the parent population of paired daily data evaluates the trade-off between significance, sample size, instrument/site differences and effect size (Figure 8).

The third problem is using the wrong test of the hypothesis that mean daily differences between instruments or sites equals zero in the case of paired t-tests; or that Tmax measured by two different instruments, or the same instrument/Stevenson screen combination at two different sites is the same.

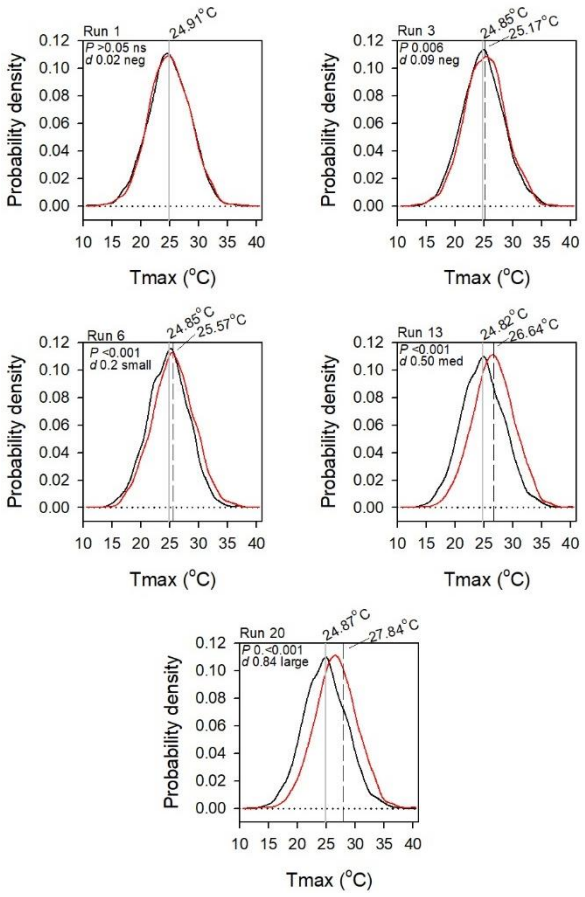

Figure 10. Probability density plots for selected scenarios used as input to Figure 9. As Tmax for Series2 increased, PDF curves became displaced; however, clear differences did not emerge until about Run 13 when *Cohens d* =0.5 (medium effect size).

As the paired t-test controls for *within subject* variation, and instruments within the same screen, or instruments taking measurements at different sites cannot sample the same parcels of air 100% of the time, even though observations occur on the same day, the paired t-test is inappropriate under the circumstances.

Conclusions

At the outset of undertaking comparisons between instruments housed in the same or different Stevenson screens or between different sites, graphical and statistical tests should be used to evaluate parallel datasets and guide subsequent analyses. The effect size statistic (*Cohens d*) provides an initial indication whether a difference between dataset means (ΔTmax/standard deviation) is likely to be meaningful in the overall scheme of things. If due to an excessive number of paired observations differences are found to be *significant* or *highly significant*, *Cohens d* less than 0.2 indicates the effect size is likely to be negligible.

T-tests are not suitable for comparing differences in closely spaced time-series unless steps are taken to identify and mitigate autocorrelation which results from embedded cycles and possible trends related to weather or other uncontrolled factors. Cycles may be removed from daily data by subtracting day-of-year (1-366) averages from respective day-of-year data to derive de-seasoned anomalies. Anomalies may also be differenced to highlight discontinuities that may be related to a third factor such as changed instrument responses.

As the paired t-test explicitly presumes all variation in the response is attributable to subjects, and meteorological instruments held within Stevenson screens are unable to sample the same parcels of air 100% of the time, even if its use is valid (for example by sub-sampling), the Type1 error-rate (declaring a difference to be significant when it is not) is considerably higher for the same data compared with the un-paired or two-samples t-test.

Dr Bill Johnston

22 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 and PAST, he does not claim to be a statistician.