Statistical concepts in climate research
(some misuses of statistics in climatology)

Slava Kharin
Canadian Centre for Climate Modelling and Analysis,
Environment Canada,
Victoria, British Columbia, Canada

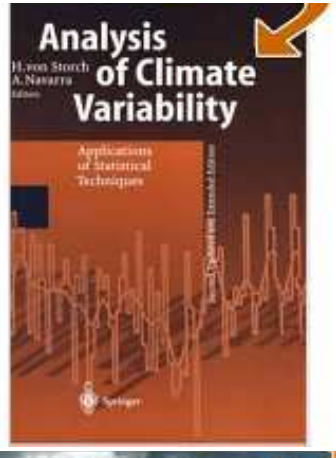
Banff Summer School, May 2008

Statistical concepts in climate research

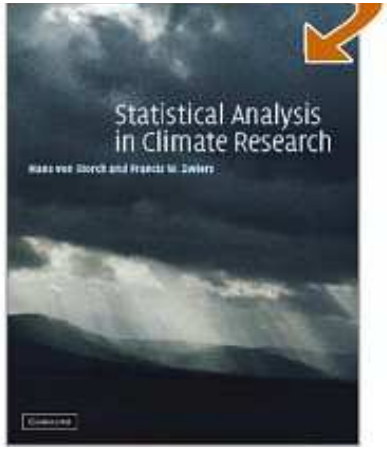
- Lecture I - Misuses of statistics in climate research
  - testing hypotheses suggested by the data;
  - serial correlation;
  - using statistical recipes as "black-box" tools;
- Lecture II - Hypothesis testing
  - Type I and Type II errors, significance, power, etc
  - historical developments and controversy around classical statistical significance test and its interpretation.
- Lecture III - Basics of Bayesian statistics.
  - Introduction to Bayesian statistics.
  - Bayesian climate change assessment.

References

Analysis of Climate Variability: Applications of Statistical Techniques, by H. von Storch and A. Navarra (Eds.)
(proceedings of an Autumn School on Elba, 1993)



Statistical Analysis in Climate Research, H. von Storch and F. W. Zwiers, 1999, Cambridge.



Misuses of Statistical Analysis in Climate Research (von Storch 1993)

- Obsession with *statistical recipes*, in particular, *hypothesis testing* – "Pavlov's dog" reaction to any hypothesis derived from data, demanding statistical significance test;
- Use of statistical techniques as a "black-box", or cook-book recipe (standard example is disregard of serial correlation).
- Misunderstanding or misinterpreting the names (e.g. *decorrelation time*, *p*-values as probability of hypotheses)
- Use of sophisticated techniques. . . There is sometimes unwarranted expectation of miracle-like results from very advanced techniques.

Statistical analysis in climate research

- There is basically one *observational* record in climate research.
  - It is hardly possible to perform real *independent experiments*. Classical hypothesis testing is "frequentist" in nature (repeated sampling).
  - Dynamical models can create independent data, subject to model deficiencies.
- Data are inter-related, both in space and time.
  - This is useful for reconstruction of the space-time state of the climate system from a limited set of observations.
  - But the basic premise of many standard statistical tests is violated.

Two types of statistical data analysis (Tukey 1977)

- *Exploratory analysis (EA)* is the art of extracting information from a data set.
  The objectives of *EA* are:
  - to suggest and formulate *hypotheses* about the workings of the climate system;
  - to assess *assumptions* on which statistical inference (i.e., estimation and hypothesis testing) will be based;
  - to support selection of statistical tools and techniques;
  - provide a basis for further data collection (e.g., through numerical experiments).
- *Confirmatory analysis (CA)* is then performed to confirm *independently* the hypotheses.

Climatology as a one-experiment science

- Climatology is a one-experiment science. There is basically one observational record in climate.
- Climate researchers essentially look at the same record to build and test their hypotheses.
- The processes of building and testing hypotheses are hardly separable.
- Only dynamical models can provide independent datasets but subject to model limitations to describe the real world. But even dynamical models (GCMs,CGCMs,ESMs) are "tuned" to the available observational record.
- This leads to *testing hypotheses suggested by the data* using the same data.

Mexican Hat

Mexican Hat is a unique combination of vertically arranged stones in the desert at the border of Utah and Arizona (google images).

Is the Mexican Hat man-made? (von Storch & Zwiers 1999)

- The null hypothesis: "The Mexican Hat is of natural origin"
- Form a statistic:

$$t(x) = \begin{cases} 1 & \text{if } x \text{ forms a Mexican Hat} \\ 0 & \text{otherwise} \end{cases}$$

  where x is any pile of rocks.
- We need the probability distribution of $t(x)$ under null hypothesis: collect samples of x, say, $n = 10^6$.
- Mexican Hat is famous for good reasons – there is only one $t(p) = 1$.
- Estimate the probability distribution of $t(p)$:

$$prob(t(x) = k) = \begin{cases} 10^{-6} & \text{for } k = 1 \\ 1 - 10^{-6} & \text{for } k = 0 \end{cases}$$

- We reject the null hypothesis at a small significance level.

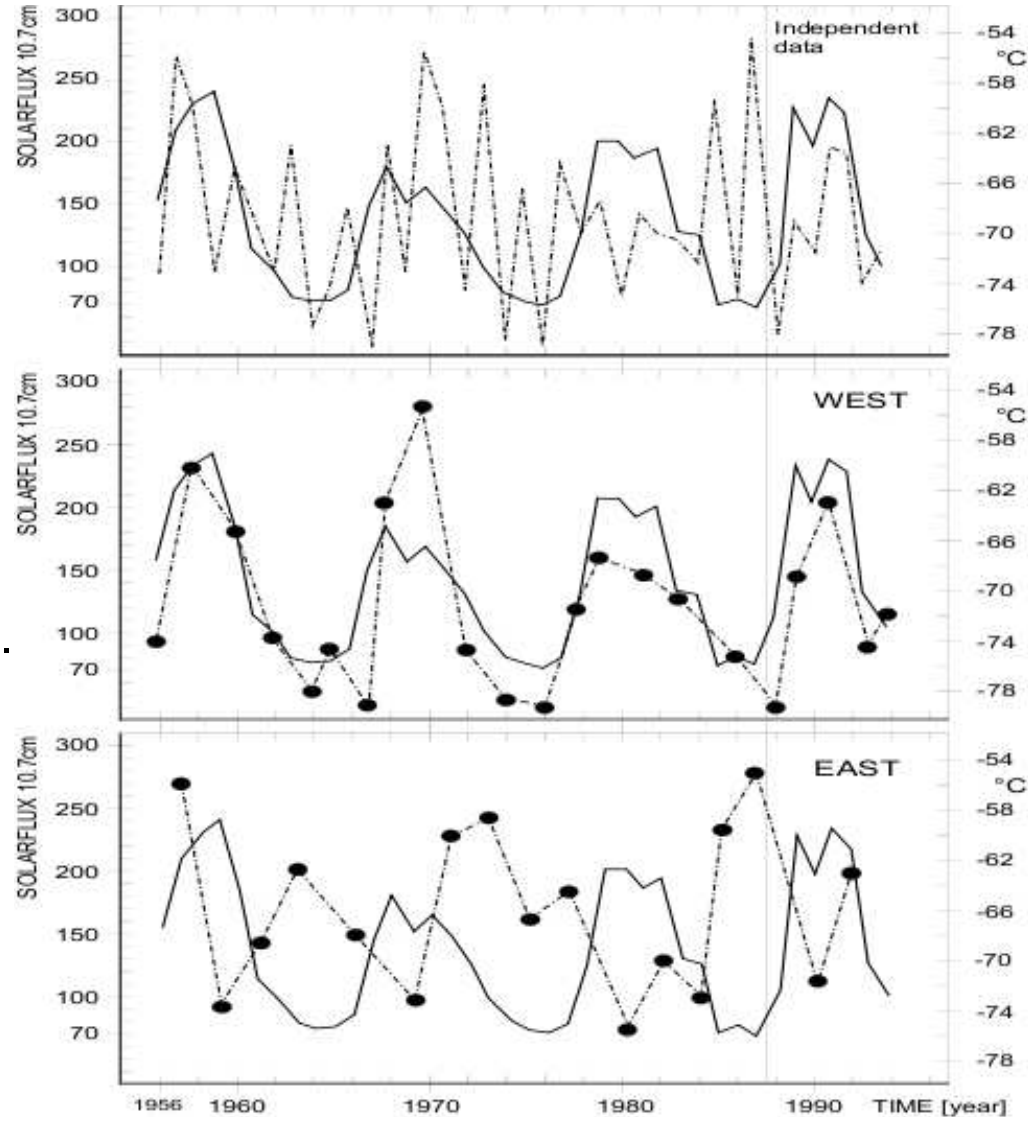
Where is the logic flaw in the Mexican Hat paradox?

Fundamental problem: the null hypothesis is not *independent* of the data which are used to conduct the test.

- We know *a priori* that the Mexican Hat is a very rare event. The fact that we didn't find any other combination like that *cannot* be used as *independent* evidence against its natural origin.
- The same trick can be used to "prove" that any rare event is "non-natural" (and therefore we are tricked into searching for external causes)

Labitzke and van Loon, 1988

- K.Labitzke studied the relationship between solar activity and the stratospheric relationship at the North Pole.
- there was no obvious relationship if all years were considered.
- She *saw* that there was *positive* correlation during WQBO and *negative* correlation during EQBO.



Testing hypotheses suggested by the data

The limitation of a "single" observational record has a severe consequence:

- Many people 'screen' the observational record for *rare* and *unusual* events.
- Typically, only "unusual" results are eventually published in literature (*publication bias*).
- Presumably, some of these "unusual" events are nevertheless "usual".
- Although we can compare these unusual events with all other events in the record, they cannot be contested with a statistical test since *independent* data are not available.

Testing hypotheses suggested by the data

No statistical test, regardless of its power, can overcome this problem!

Possible solutions:

- extend record backwards;
- use suitably designed experiments with GCMs;
- postpone testing until new data become available.

Testing hypotheses suggested by the data

No statistical test, regardless of its power, can overcome this problem!

Possible solutions:

- extend record backwards;
- use suitably designed experiments with GCMs;
- postpone testing until new data become available.

Neglecting Serial Correlation

- *Problem:* Most standard statistical techniques are derived with explicit need for statistically independent data. However, almost all climatic data are somehow correlated in time.
- *Consequence:* Statistical tests becomes too liberal, that is, the "no-signal" hypothesis is incorrectly rejected more often than it is implied by the significance level when it is true.
- The results of a statistical tests depends *strongly* on the serial correlation.
- Even if serial correlation is not neglected, its effects are not always correctly accounted for in statistical tests.

Comparison of means test (Zwiers & von Storch 1994)

There are two main assumptions in the standard t-test:

- *sampling* assumption – observations are statistically independent;
- *distributional* assumption – observations are normally distributed;

t-statistic for one sample X:

$$t = \frac{\overline{X} - \mu_0}{S_X/\sqrt{n}} = \frac{D}{S_D}$$

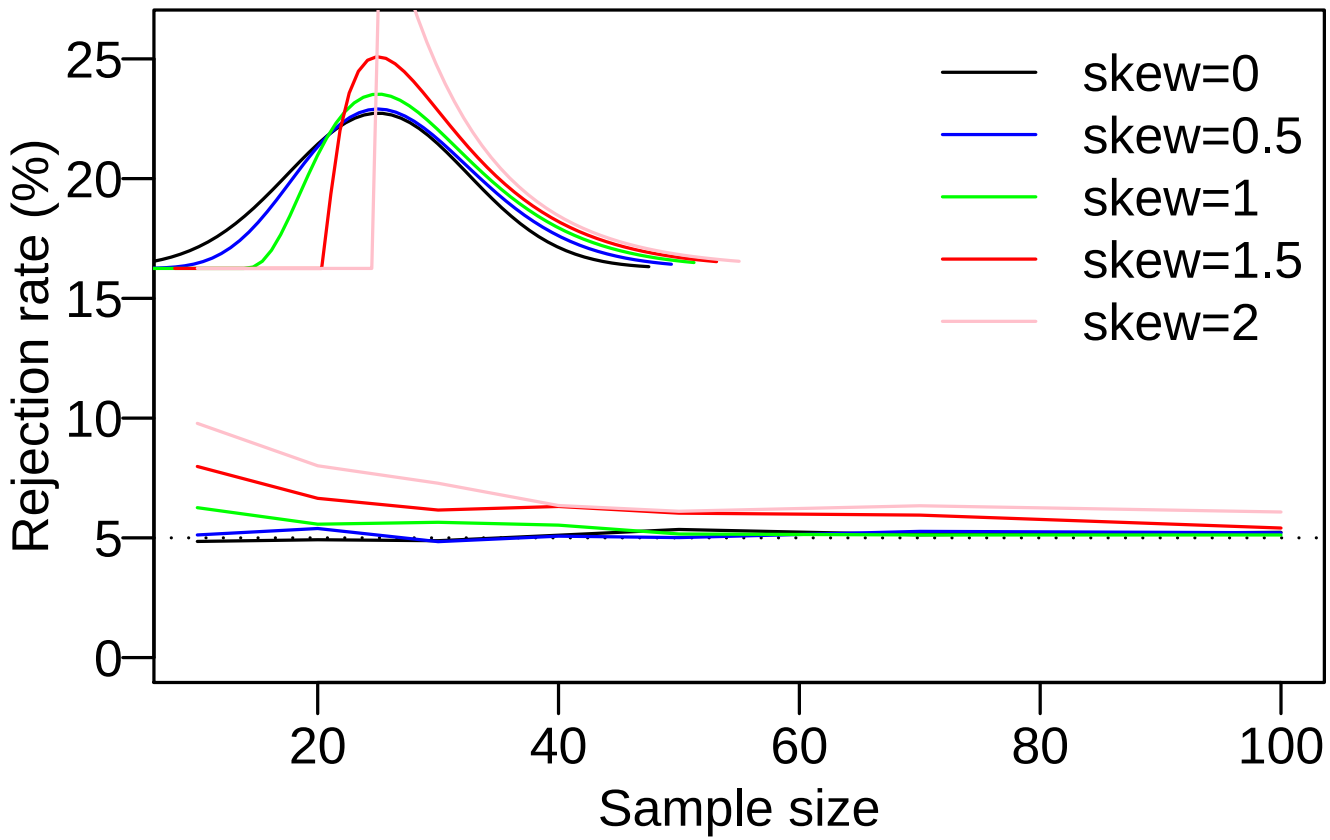
where S_X is the sample standard deviation.
The t-statistic has Student's t distribution with $n - 1$ d.f.

Comparison of means test: violation of assumptions

How sensitive is the *t*-test to departures from the sampling and distributional assumptions?

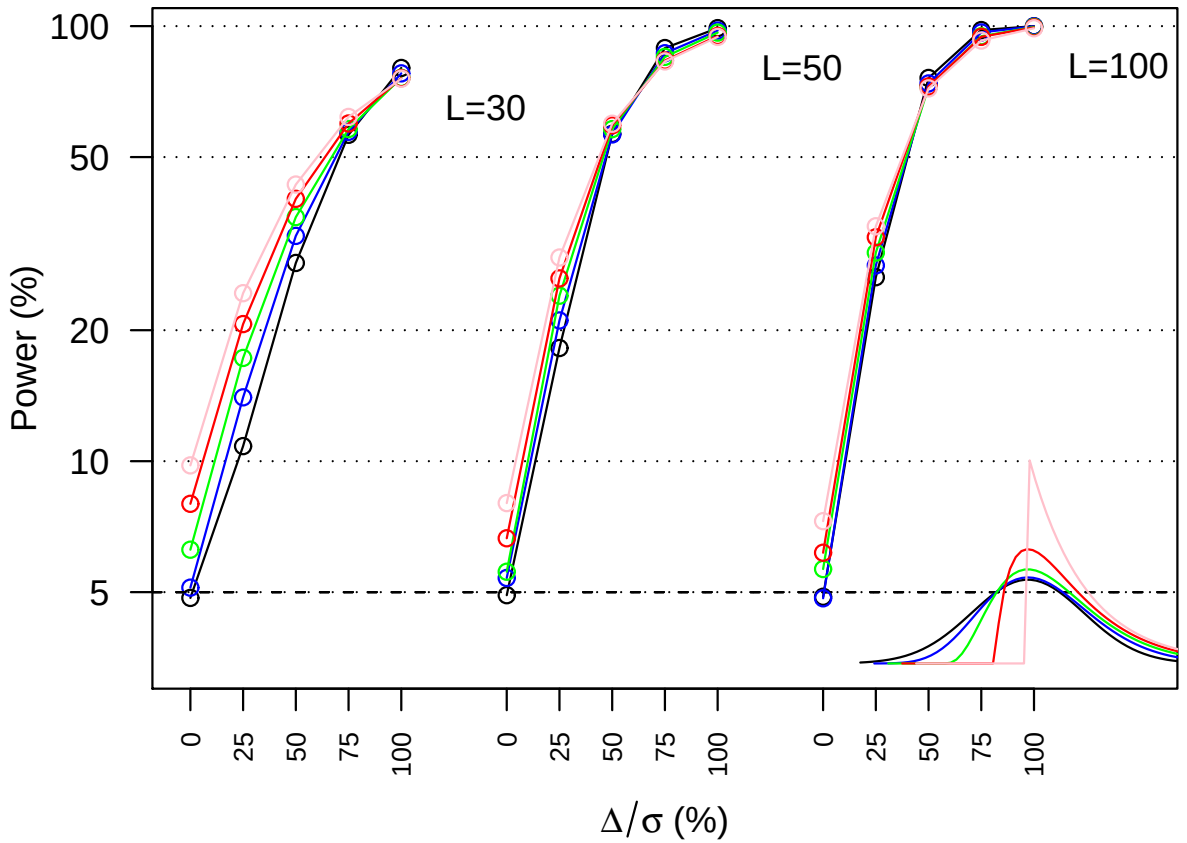
- the *t*-test is *moderately robust* against departures from the normal distribution, especially when large samples are available, but it depends on the nature of deviations from the normality.
- the *t*-test is *strongly* affected by serial correlation, both for small or large samples.

Comparison of means test – Type I errors: departures from the normality (Pearson distribution family)

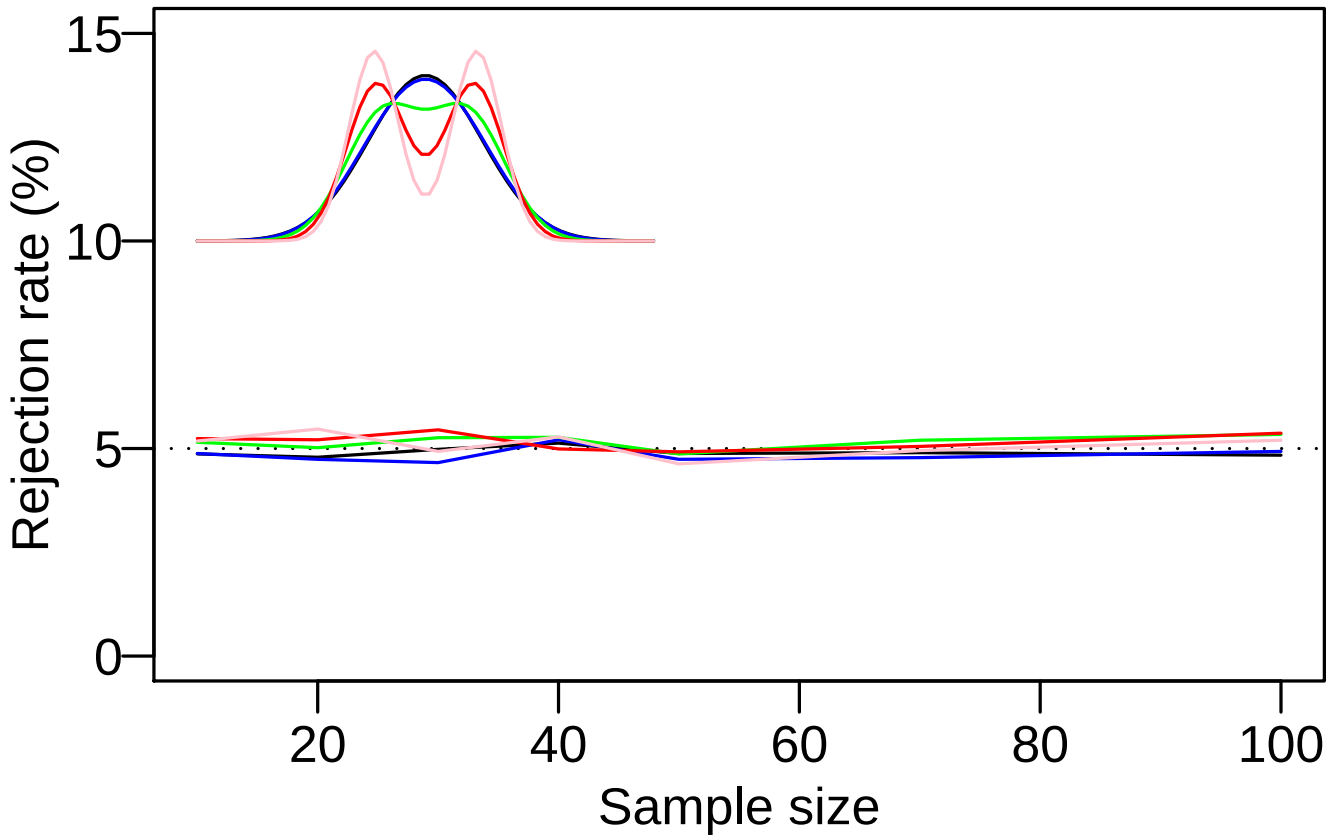


The *t* test is slightly too *liberal* when distributions are skewed but not excessively so.

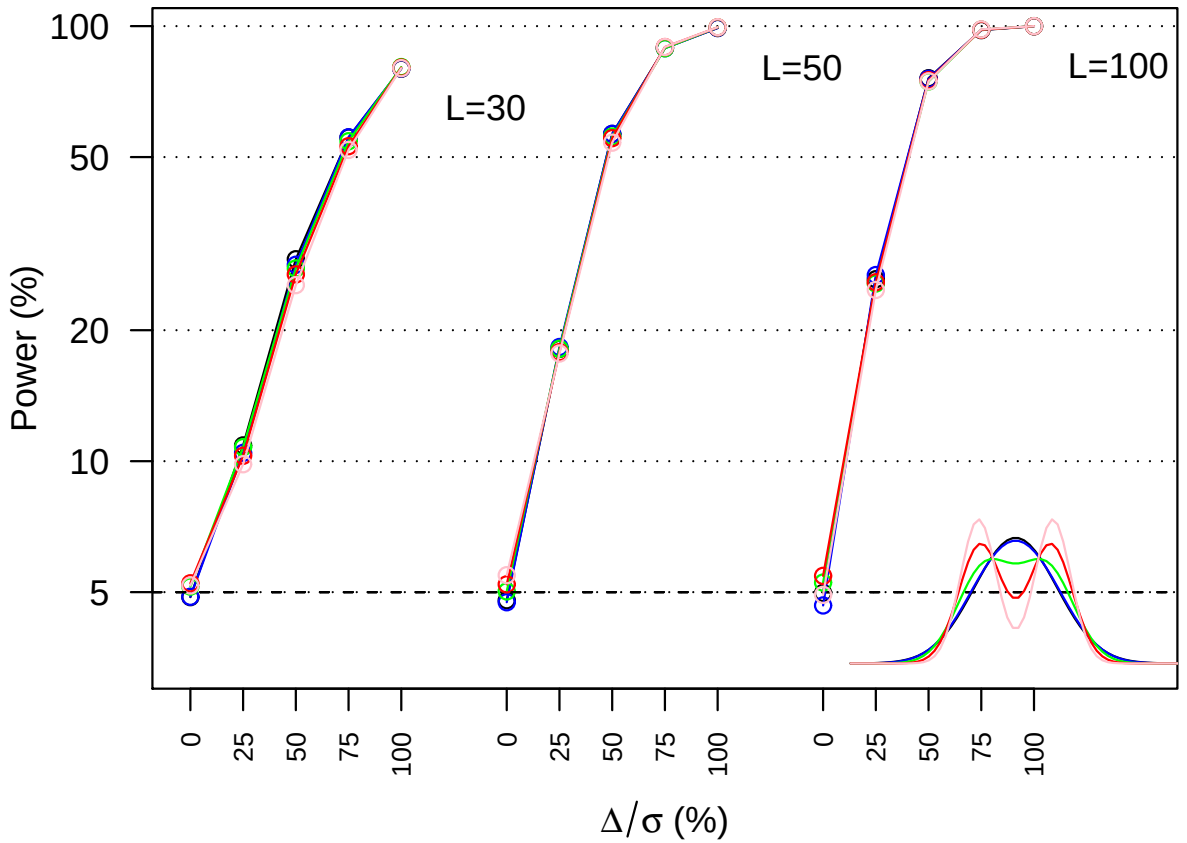
Comparison of means test – Power: departures from the normality (Pearson distribution family)



Comparison of means test – Type I errors: departures from the normality (bimodal)



Comparison of means test – Power: departures from the normality (bimodal)



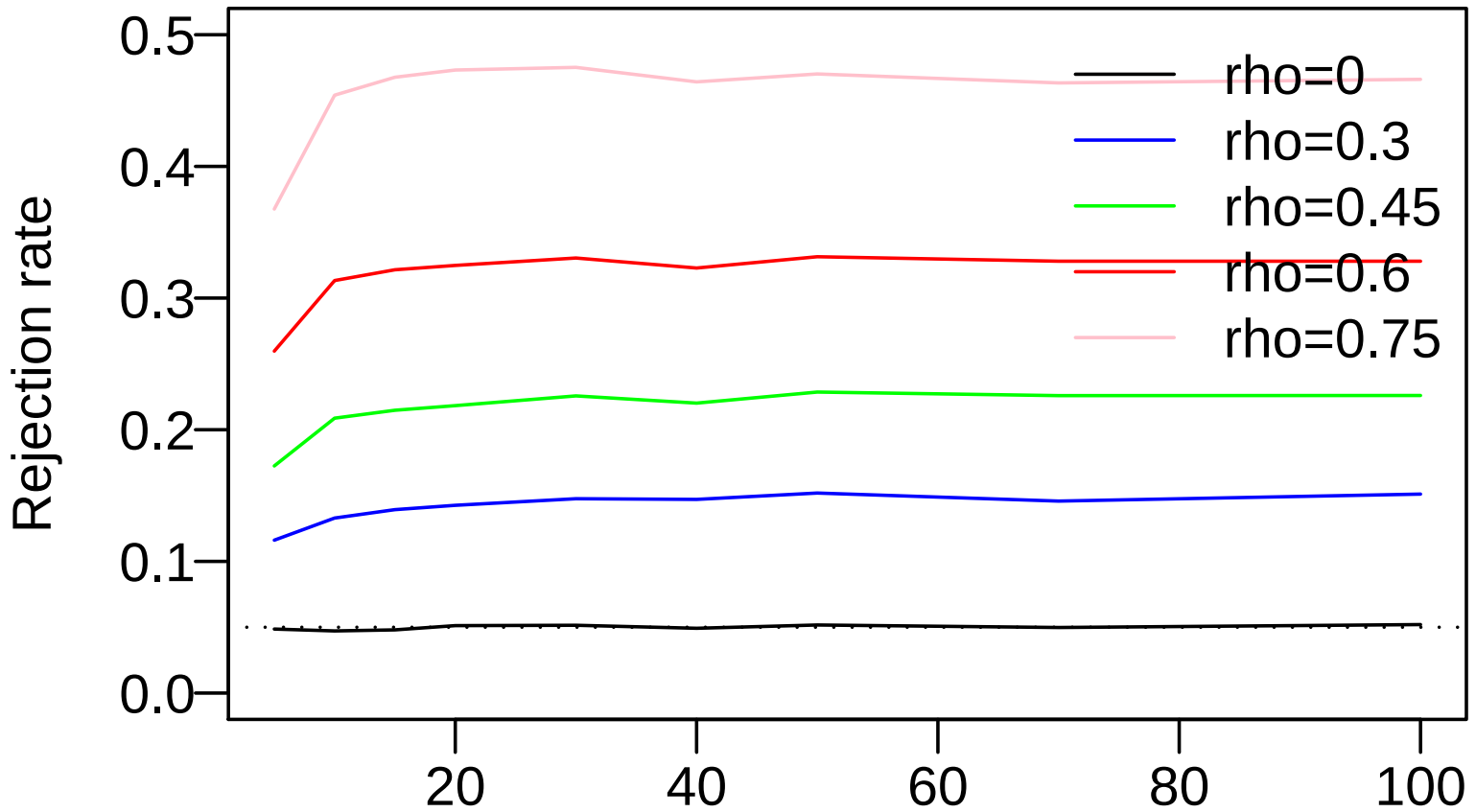
Comparison of means test: departures from the independence

We assume that a climate process can be approximately represented by a auto-regressive first order (AR1) process:

$$X_t - \mu_X = \rho_1 (X_{t-1} - \mu_X) + \epsilon_t$$

where ϵ_t is a Gaussian white noise process, ρ_1 is the lag-1 correlation coefficient.
(ENSO, QBO, or MJO are suspect)

Comparison of means test: departures from the independence



Subsampling the data

- If a data record is long, one can try to form subsamples, separated by some time interval.

$$t = \frac{\overline{x}^* - \mu}{S_X^*/\sqrt{n^*}} = \frac{D^*}{S_D^*}$$

- if one is interested in the differences of means in a particular season (DJF), it is often more prudent to use seasonal means, rather than monthly means.
- Choice of sampling interval is not trivial (either from physically considerations, or from inspecting autocorrelation function).
- throwing away much of data.

The equivalent sample size, ESS

The main reason why the t test is too liberal for serially correlated data is that the denominator *underestimate* the sampling standard errors of the numerator:

- The concept of "*effective or equivalent sample size*" is used to correct this problem (Thiebaux and Zwiers (1984) discuss interpretation and estimation of ESS).
- If $\{X_1, \ldots, X_n\}$ are n i.i.d. random variables then

$$Var(\overline{X}) = \frac{1}{n} Var(X)$$

- If $\{X_1, \ldots, X_n\}$ are serially correlated then

$$Var(\overline{X}) = \frac{1}{n_e} Var(X)$$

$$n_e = \frac{n}{1 + 2\sum_{\tau=1}^{n-1}(1 - \frac{\tau}{n})\rho(\tau)} \approx n\frac{1-\rho_1}{1+\rho_1}, n \gg 1,$$

$\rho(\tau)$ is the autocorrelation function.

The equivalent sample size

- *Interpretation:* samples of independent observations of size n_e contain as much information about the *difference of means* as samples of serially correlated observations of size n.
- *Important:* ESS is not uniquely defined.

$$T = D/S_D$$

The expression for n_e depends on the definition of D. If our interest is something else (e.g., variance) the definition of information, and therefore of the ESS, changes.

- Therefore, it is incorrect to interpret n_e as the number of effectively independent observations in the absolute sense.

t-test with the equivalent sample size

- When samples are *sufficiently large* then

$$t = \frac{\overline{X} - \mu_0}{S_X / \sqrt{n_e}}$$

  has approximately a standard normal distribution $\mathcal{N}(0, 1)$.
- When samples are small it is often *assumed* that t behaves as Student's t statistic with $(2n_e - 2)$ d.f. Correct?
- Wrong! This assumption is not correct for small samples.

t-test with the equivalent sample size

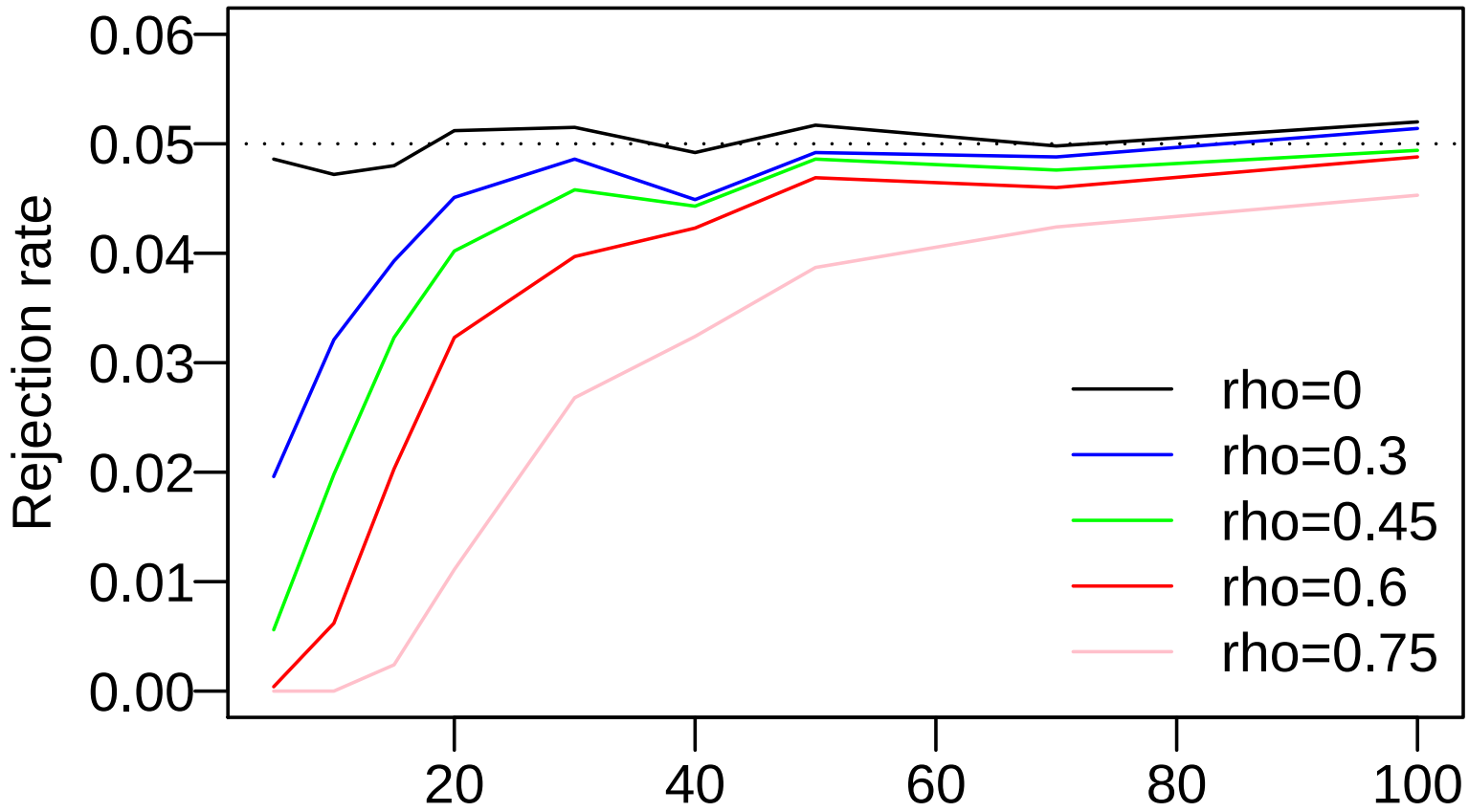
- When samples are *sufficiently large* then

$$t = \frac{\overline{X} - \mu_0}{S_X / \sqrt{n_e}}$$

  has approximately a standard normal distribution $\mathcal{N}(0, 1)$.
- When samples are small it is often *assumed* that t behaves as Student's t statistic with $(2n_e - 2)$ d.f. Correct?
- Wrong! This assumption is not correct for small samples.

t-test with the *known* equivalent sample size

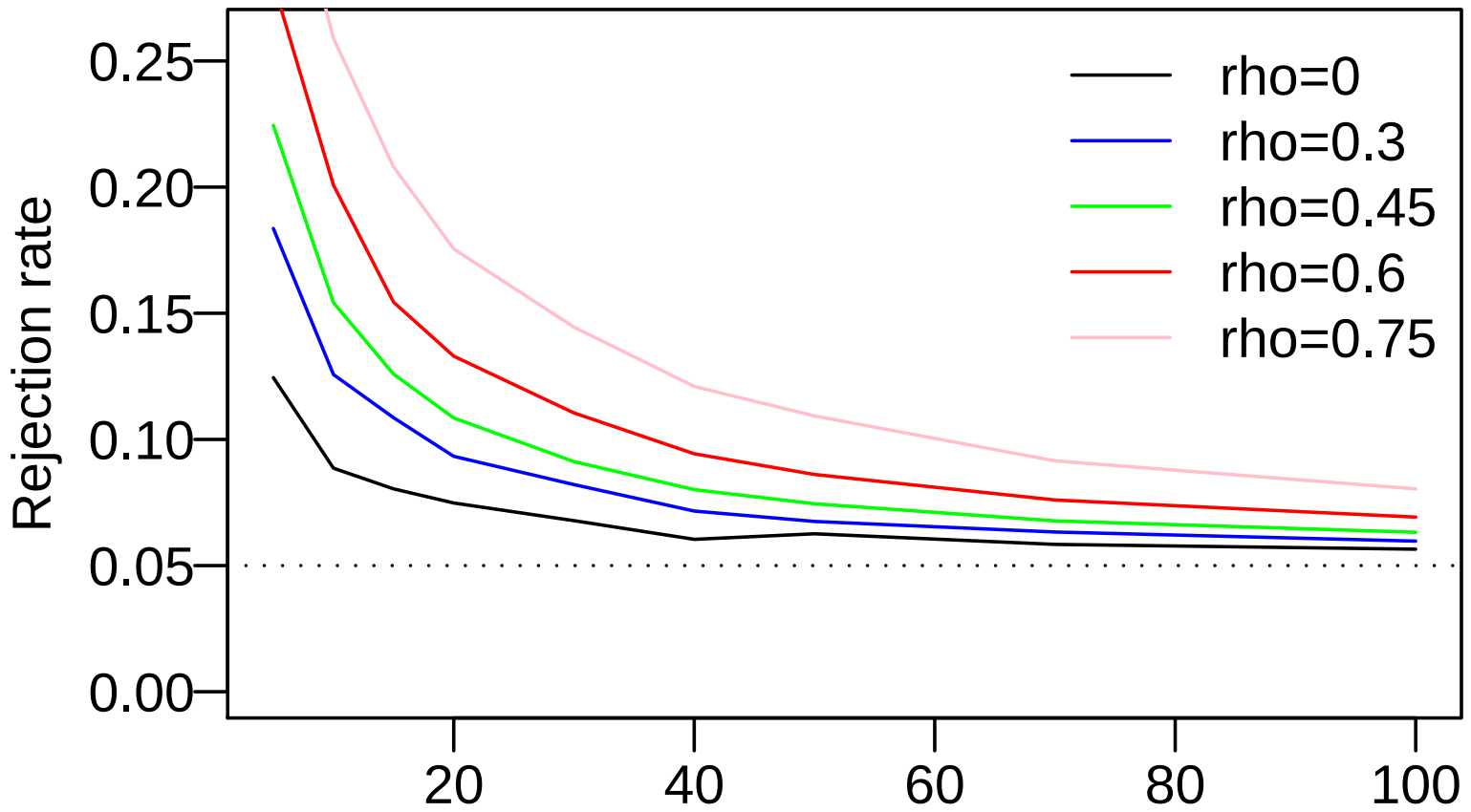
The actual rejection rate of the t test with known n_e (adapted after Zwiers & von Storch 1994)



The t test is too *conservative* for small sample sizes, and therefore, wrong.

t-test with the *estimated* equivalent sample size

The actual rejection rate of the t test with n_e *estimated* from the data



The t test is too *liberal* for small sample sizes, and therefore, wrong.

t-test with the "table look-up"

Zwiers and Storch (1994) recommend:

- t test *should not* be used with equivalent sample sizes smaller than 30 (note that 30 is for the ESS, not sample size).
- If ESS$<$30, table look-up test should be used (2 sample test):
  - compute standard $t = \frac{\overline{X}-\overline{Y}}{\sqrt{(S_X^2+S_Y^2)/n}}$
  - compute the pooled lag-1 correlation coefficient

$$\rho_1 = \frac{\sum_{i=2}^{n} x'_i x'_{i-1} + \sum_{i=2}^{n} y'_i y'_{i-1}}{\sum_{i=1}^{n} y'^2 + \sum_{i=1}^{n} y'^2_i}$$

- Use the look-up tables for n and ρ_1 to determine critical values.

Example of using the ESS in literature

Effects of the El Niño–Southern Oscillation and the Quasi-Biennial Oscillation on polar temperatures in the stratosphere

C. I. Garfinkel[1] and D. L. Hartmann[1]

Received 29 January 2007; revised 20 June 2007; accepted 3 July 2007; published 13 October 2007.

[1] Reanalysis data are used to study the effects of the Quasi-Biennial Oscillation (QBO) and the El Niño–Southern Oscillation (ENSO) on the stratosphere. During the boreal winter in the Arctic, Warm ENSO (WENSO) months are found to be significantly warmer and Cold ENSO (CENSO) months significantly colder than climatology. The QBO is also found to have a large effect on the Arctic stratosphere during the late fall/early winter; Westerly QBO (WQBO) poles are colder, and Easterly QBO (EQBO) poles are warmer. In the first half of the 50 years of interest, WENSO and EQBO have had a tendency to be correlated in time, and thus their signals are difficult to disentangle. In order to isolate each effect from the other, composites are taken of QBO months under near-neutral ENSO conditions, which show a clear effect in late fall/early winter. Because of the bimodality of QBO, producing a meaningful composite of ENSO months under near-neutral QBO is difficult, as the number of available months is quite small. To distinguish ENSO from QBO and to further study the QBO, we compare composites of months with four different combinations of QBO and ENSO anomalies, which confirms that ENSO does have a significant effect on the polar vortex. These groupings are also studied after removing the 2 years following one of the three major volcanic eruptions during the 50 years of data and during the post-1979 satellite era only as well. These composites show distinct ENSO and QBO effects of comparable magnitude.

Example of using the ESS in literature

- Use ERA-40 reanalyses to study ENSO/QBO effects on polar temperature, e.g. in the NDFJ season ($\approx$ 200 months).
- composites for different ENSO/QBO phases are based on 10-50 months. Use monthly data to increase sample size.
- ESS is estimated as $N_e = N\frac{1-r}{1+r}$ (an approximation for N is large, bad small sample behaviour).
- normality is evaluated in MC tests assuming that monthly means are independent.
- perform a great number of statistical significance tests.

This study highlight several possible "misuses" of statistics:

- obsession with hypothesis testing (perhaps providing confidence intervals would be more preferable)
- the use of statistical techniques like a cook-book recipe (equivalent degrees of freedom)

Example of using the ESS in literature

- Use ERA-40 reanalyses to study ENSO/QBO effects on polar temperature, e.g. in the NDFJ season ($\approx$ 200 months).
- composites for different ENSO/QBO phases are based on 10-50 months. Use monthly data to increase sample size.
- ESS is estimated as $N_e = N\frac{1-r}{1+r}$ (an approximation for N is large, bad small sample behaviour).
- normality is evaluated in MC tests assuming that monthly means are independent.
- perform a great number of statistical significance tests.

This study highlight several possible "misuses" of statistics:

- obsession with hypothesis testing (perhaps providing confidence intervals would be more preferable)
- the use of statistical techniques like a cook-book recipe (equivalent degrees of freedom)

Conclusions

- *Don't be obsessed with hypothesis testing* when examining observational data
  – statistical hypothesis testing is problematic and cannot be viewed as objective and unbiased judge of the null hypothesis under these circumstances;
- *Be careful about the effects of serial correlation* on the results of statistical inference (such as estimation and hypothesis testing)
  – some 'off-the-shelf" methods and tools behave badly for small samples.

References

- Analysis of climate variability – Application of Statistical Techniques. Edited by H. von Storch and A. Navarra. Springer. 1995.
- Thiébaux, H. J., and F. W. Zwiers, 1984: The Interpretation and Estimation of Effective Sample Size. *J. Appl. Meteor.*, **5,** 800–811
- Francis W. Zwiers, and H. von Storch, 1994: Taking Serial Correlation into Account in Tests of the Mean. *J. Climate*, *8,* 336–351.
- Zhang, X., and F.W. Zwiers: 2004: Comment on "Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test". *Water Resources Research*.
- von Storch, H., and F. W. Zwiers, 1999: Statistical Analysis in Climate Research. Cambridge.