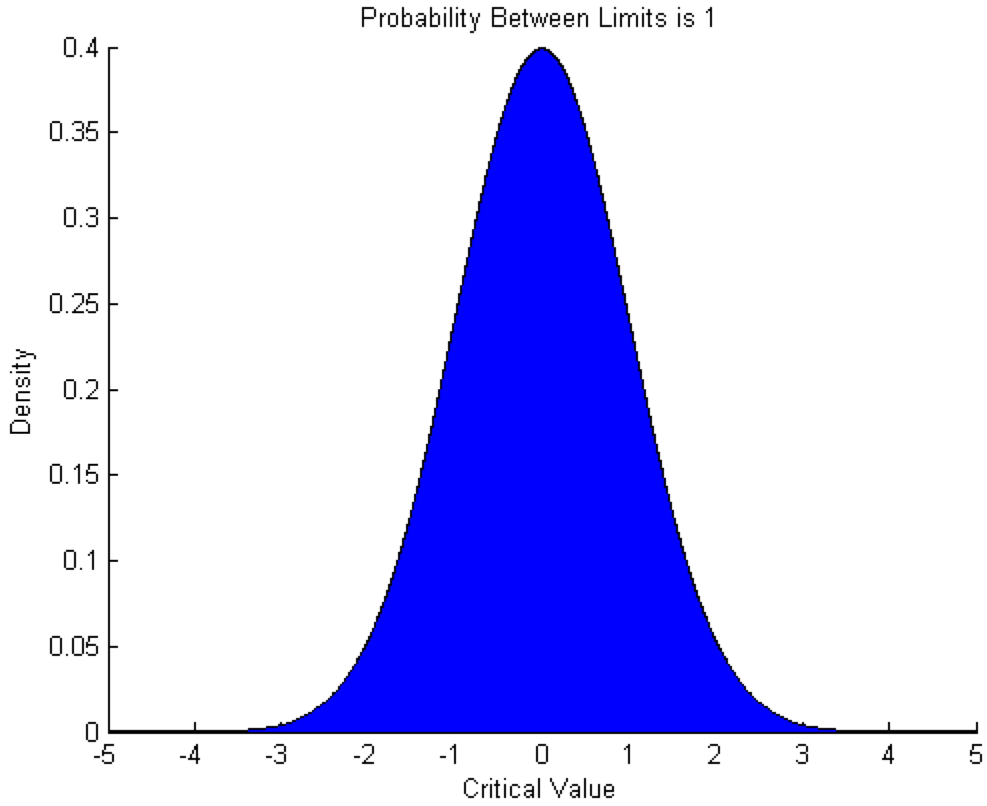
Assignments

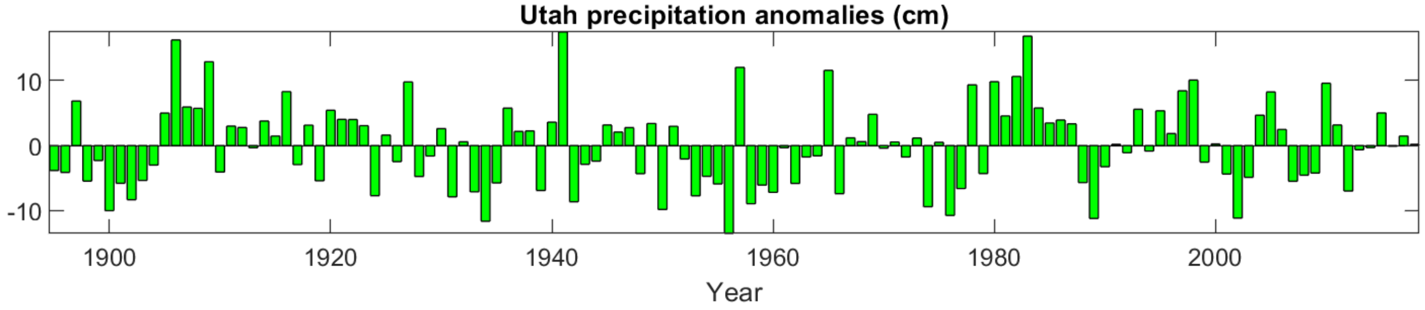
- Read Chapter 3 and 4 notes in Canvas (Chapter 5 also available)
- Read text: Text Chapter 3.1-3.6, Chapter 4
- Assignment 10 released soon

Empirical vs. Parametric Distributions

- Parameteric distributions:
  - Theoretical approach to define populations with known properties
  - Can be defined by a function with couple parameters and assumption that population composed of random events



Annual Precip in Utah



What 3 yr periods have experienced droughts?

Steps of Hypothesis Testing

- Identify a test statistic that is appropriate to the data and question at hand
  - Computed from sample data values. 3 yr sample means
- Define a null hypothesis, H_0 to be rejected
  - 3 yr sample mean 0
- Define an alternative hypothesis, H_A
  - 3 yr sample mean < 0
- Estimate the null distribution
  - Sampling distribution of the test statistic IF the null hypothesis were true
  - Making assumptions about which parametric distribution to use (Gaussian, Weibull, etc.)
  - Use sample mean of 0 and 122 yr sd of 6.6
- Compare the observed test statistics (3-yr means) to the null distribution. Either
  - Null hypothesis is rejected as too unlikely to have been true IF the test statistic fall in an improbable region of the null distribution
    - Possibility that the test statistics has that particular value in the null distribution is small
    - normspec([-1.96*6.4,1.96*6.4],0,6.4)
    - OR
    - The null hypothesis is not rejected since the test statistic falls within the values that are relatively common to the null distribution

Central Limit Theorem

sum (or mean) of a sample (6 dice) will have a Gaussian distribution even if the original distribution (one die) does not have a Gaussian distribution, especially as the sample size increases.

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

$\sigma_{\bar{x}}$ standard deviation of the sample means

σ standard deviation of the original population

n sample size

$$s_x = \sqrt{\frac{n-1}{n}}\sigma \qquad \sigma_{\bar{x}} = s_x / \sqrt{n-1}$$

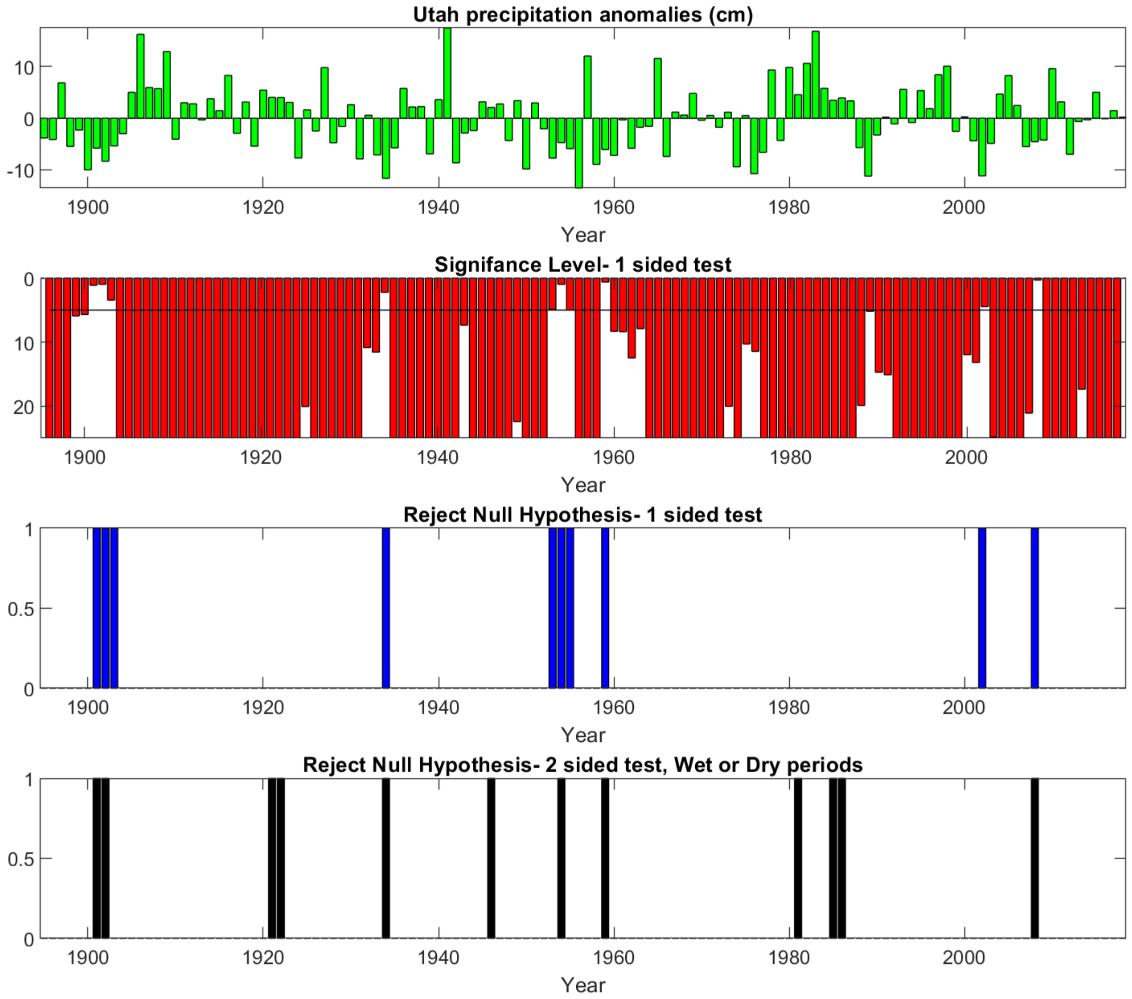
degrees of freedom: n-1, since sample can be described by the mean (1 value) plus n-1 others

Students’ t test

- $\sigma_{\overline{X}} = \frac{s_x}{\sqrt{n-1}}$
- Estimate of population variance from sample
- T value:
- Numerator: signal
- Denominator: noise
- At t gets larger, confidence in rejecting the null hypothesis (sample mean differs from population mean) gets higher
- T large IF:
  - Spread between sample and population means large
  - Degrees of freedom is large
  - Variability in sample is small

$$t = (\bar{x} - \mu)\sqrt{n-1} / s_x$$

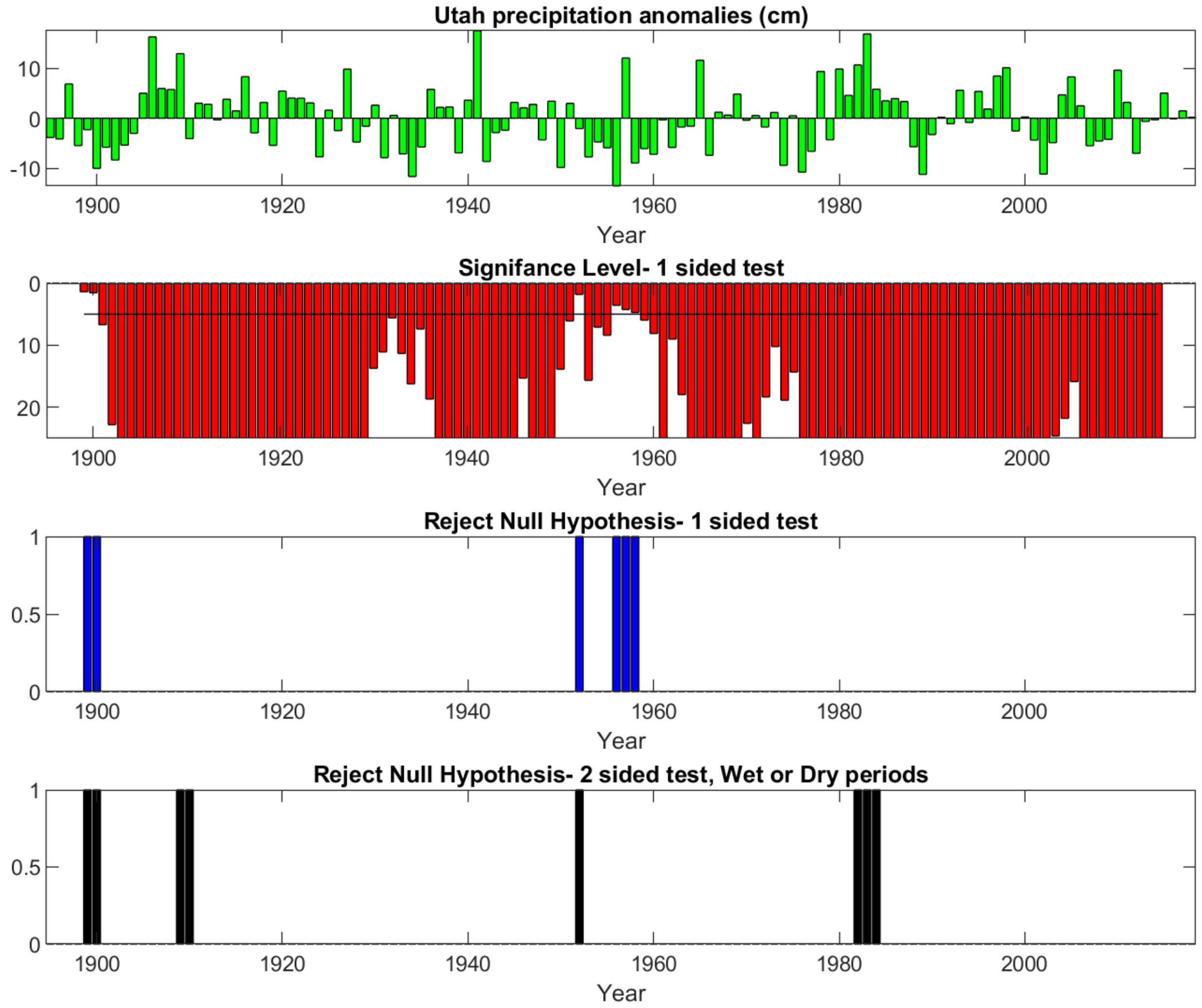
Which 3-yr samples would be considered a drought?



Left, Right, 2 Sided (both)

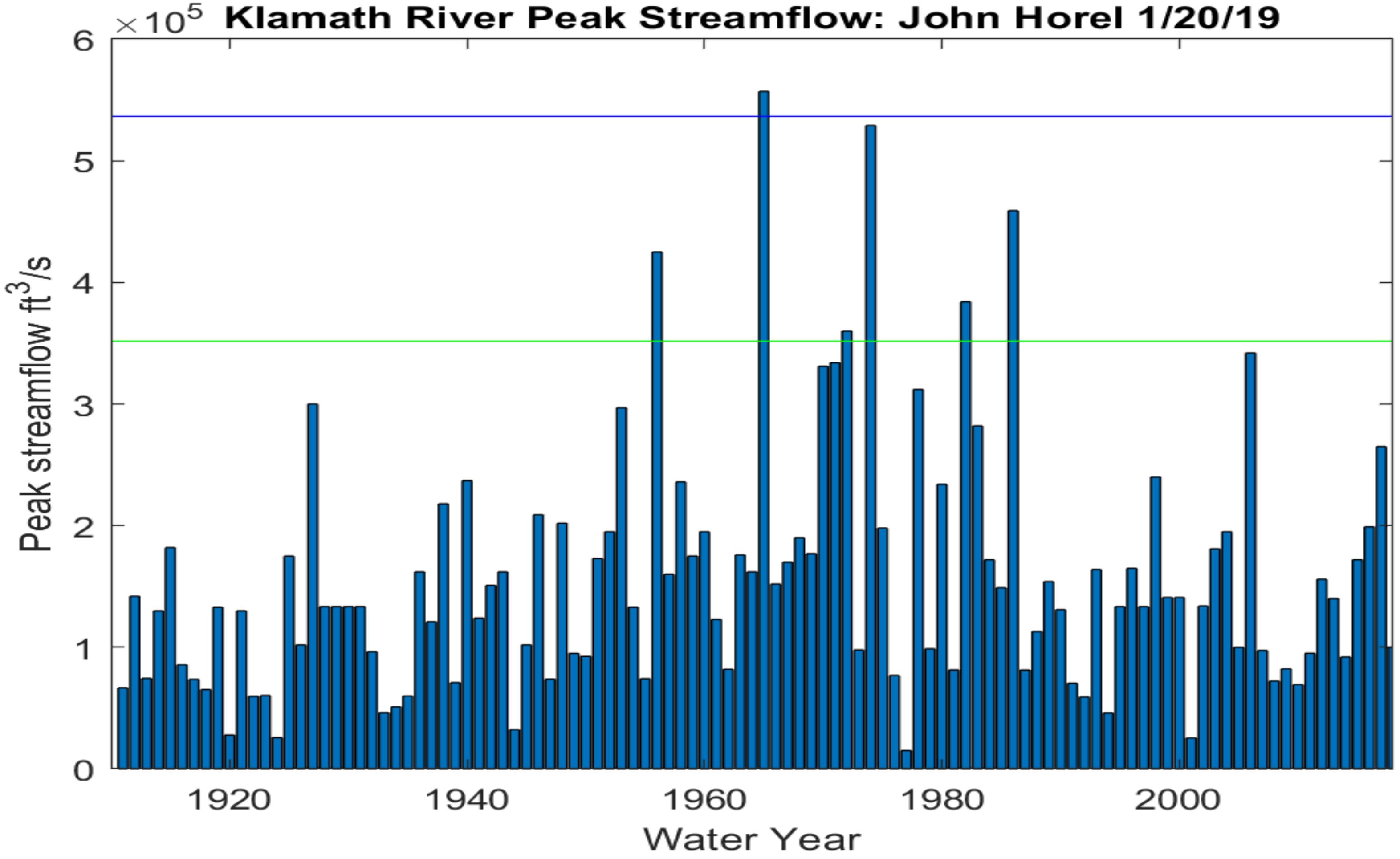
- Left- alternative hypothesis is drought
- Right- alternative hypothesis is flood
- 2 Sided- either drought or flood
- 2 sided tests are weaker and should be avoided
- [h,p,ci,stat]= ttest(valy,0,.05,'both')
- Sample value must be further from 0 (smaller p value) since α is smaller by 2 (2.5% in each tail)

Two Sided Test: Flood or Drought



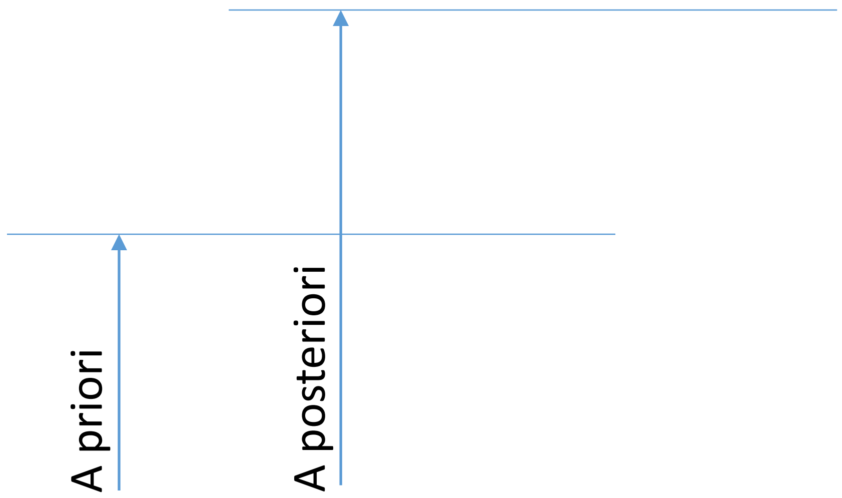
Klamath River Streamflow

- Analyzing “flow” variable

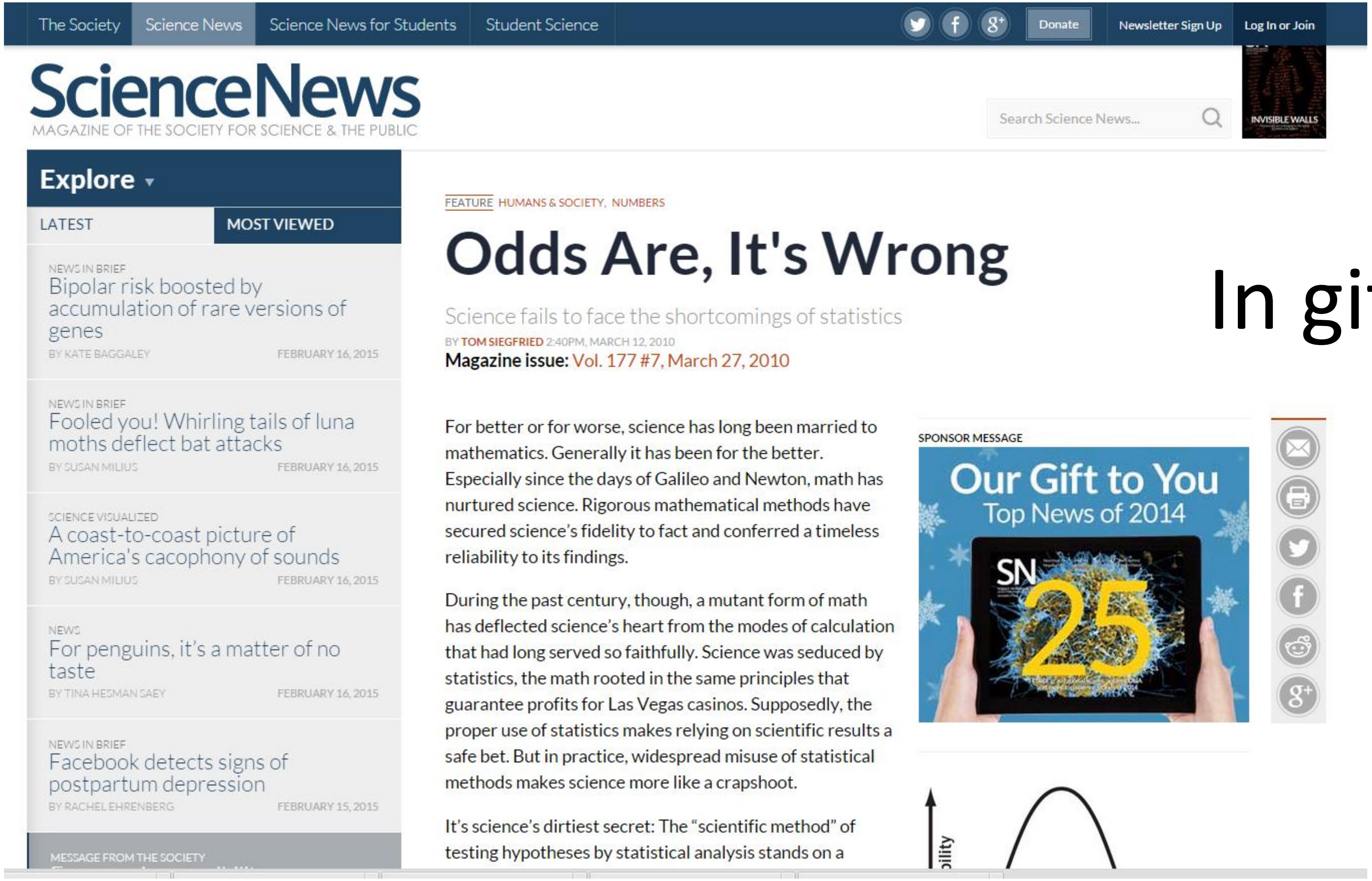


Setting the statistical bar

- a priori ("from the earlier") vs. a posteriori ("from the later")
  - a priori- can be verified independently: any conclusions drawn do not depend on a prior analysis of empirical evidence
  - a posteriori- involves some level of subjective exposure to the data from which a conclusion will be drawn



https://www.sciencenews.org/article/odds-are-its-wrong



The Society | Science News | Science News for Students | Student Science | Donate | Newsletter Sign Up | Log In or Join

ScienceNews

MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Search Science News...

Explore

LATEST | MOST VIEWED

NEWS IN BRIEF
Bipolar risk boosted by accumulation of rare versions of genes
BY KATE BAGGALEY FEBRUARY 16, 2015

NEWS IN BRIEF
Fooled you! Whirling tails of luna moths deflect bat attacks
BY SUSAN MILIUS FEBRUARY 16, 2015

SCIENCE VISUALIZED
A coast-to-coast picture of America's cacophony of sounds
BY SUSAN MILIUS FEBRUARY 16, 2015

NEWS
For penguins, it's a matter of no taste
BY TINA HESMAN SAEY FEBRUARY 16, 2015

NEWS IN BRIEF
Facebook detects signs of postpartum depression
BY RACHEL EHRENBERG FEBRUARY 15, 2015

MESSAGE FROM THE SOCIETY

FEATURE HUMANS & SOCIETY, NUMBERS

Odds Are, It's Wrong

Science fails to face the shortcomings of statistics

BY TOM SIEGFRIED 2:40PM, MARCH 12, 2010

Magazine issue: Vol. 177 #7, March 27, 2010

For better or for worse, science has long been married to mathematics. Generally it has been for the better. Especially since the days of Galileo and Newton, math has nurtured science. Rigorous mathematical methods have secured science's fidelity to fact and conferred a timeless reliability to its findings.

During the past century, though, a mutant form of math has deflected science's heart from the modes of calculation that had long served so faithfully. Science was seduced by statistics, the math rooted in the same principles that guarantee profits for Las Vegas casinos. Supposedly, the proper use of statistics makes relying on scientific results a safe bet. But in practice, widespread misuse of statistical methods makes science more like a crapshoot.

It's science's dirtiest secret: The "scientific method" of testing hypotheses by statistical analysis stands on a

In github!

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3819157/

Journal of Psychiatry & Neuroscience
Submit a Manuscript | Email Alerts | About JPN

J Psychiatry Neurosci. 2013 Nov; 38(6): 423–426.
doi: 10.1503/jpn.130002
PMCID: PMC3819157

Elevated incidence of suicide in people living at altitude, smokers and patients with chronic obstructive pulmonary disease and asthma: possible role of hypoxia causing decreased serotonin synthesis

Simon N. Young, PhD

Abstract

Go to:

Recent research indicates that suicide rates are elevated in those living at higher altitudes in both the United States and South Korea. A possible mechanism that was proposed is metabolic stress associated with hypoxia. This commentary discusses these results, and also the association between elevated suicide rates and other conditions associated with hypoxia (smoking, chronic obstructive pulmonary disease and asthma). Tryptophan hydroxylase may not normally be saturated with oxygen, so mild hypoxia would decrease serotonin synthesis. Low brain serotonin is known to be associated with suicide. Thus, the commentary proposes and discusses the hypothesis that decreased brain serotonin synthesis associated with hypoxia is a mechanism that may contribute to suicide in conditions causing hypoxia. Finally the commentary proposes various studies that could test aspects of this hypothesis.

Home » News

Salt Lake City 52 °
Traffic / Ski Report

The Salt Lake Tribune
WWW.SLTRIB.COM FEB 15, 2015

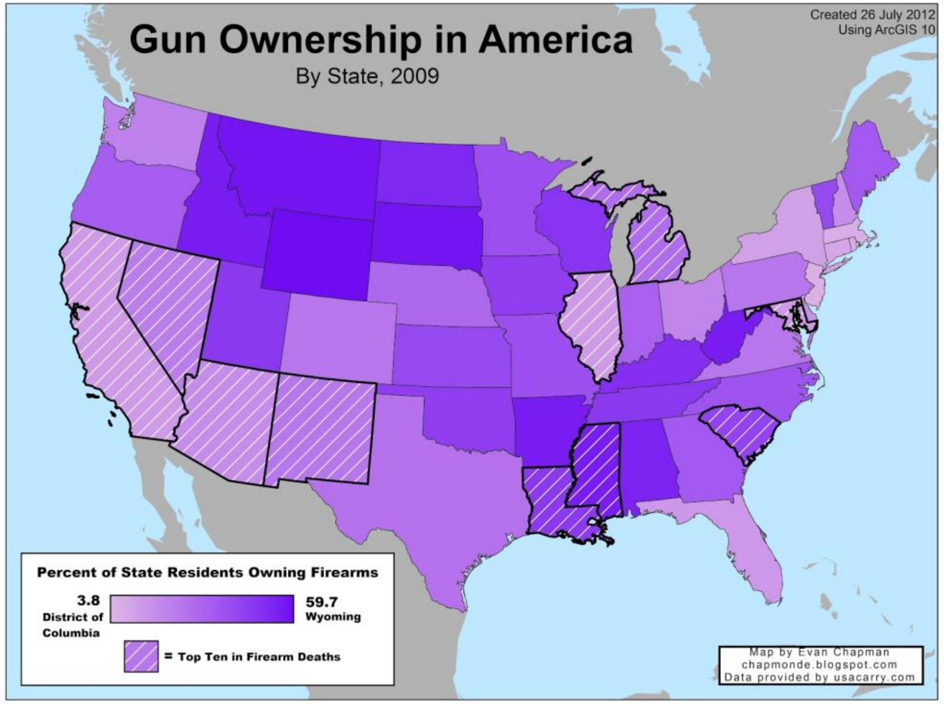
U. study: Middle-aged men at highest risk of suicide after breathing bad air

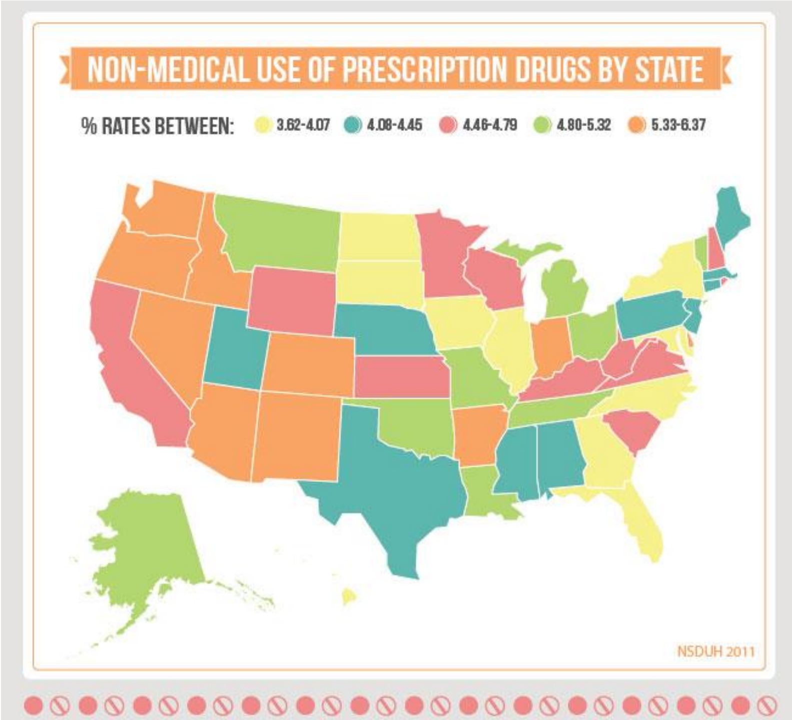
By KRISTEN MOULTON | The Salt Lake Tribune CONNECT

First Published Feb 12 2015 02:03PM • Last Updated Feb 12 2015 10:04 pm

- We examined the associations between short-term exposure to nitrogen dioxide, particulate matter, and sulfur dioxide and completed suicide in Salt Lake County, Utah (n = 1,546) from 2000 to 2010. We used a time-stratified case-crossover design to estimate adjusted odds ratios for the relationship between suicide and exposure to air pollutants on the day of the suicide and during the days preceding the suicide. We observed maximum heightened odds of suicide associated with interquartile-range increases in nitrogen dioxide during cumulative lag 3 (average of the 3 days preceding suicide; odds ratio (OR) = 1.20) and fine particulate matter (diameter ≤2.5 μm) on lag day 2 (day 2 before suicide; OR = 1.05). Following stratification by season, an increased suicide risk was associated with exposure to nitrogen dioxide during the spring/fall transition period (OR = 1.35) and fine particulate matter in the spring (OR = 1.28) during cumulative lag 3.

Contributing factors?





ESSAY

AWARENESS OF BOTH TYPE 1 AND 2 ERRORS IN CLIMATE SCIENCE AND ASSESSMENT

BY WILLIAM R. L. ANDEREGG, ELIZABETH S. CALLAWAY, MAXWELL T. BOYKOFF, GARY YOHE, AND TERRY L. ROOT

Climate science and assessment sometimes focus too strongly on avoiding false-positive errors, when false-negative errors may be just as important.

https://journals.ametsoc.org/doi/full/10.1175/BAMS-D-13-00115.1

Situation / Decision	**Null hypothesis is TRUE**	**Null hypothesis is FALSE**
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Fig. 1. Graphical representation of type 1 and type 2 errors.

Type 1 errors are a false positive: a researcher states that a specific relationship exists when in fact it does not. Type 1 errors are typically avoided in hypothesis testing by determining whether a p value, roughly the probability that a result could be obtained by chance alone, falls below a predetermined threshold.

Type 2 errors are the reverse: a null hypothesis would not be rejected despite being false- A scientist says no relationship exists when, in fact, one exists

- Scientific method tends to be averse to type 1 errors and risks committing type 2 errors
- Tendency to err on the side of least drama, scientific reticence
- High consequence and tails of the distribution of climate impacts, where experts may disagree on likelihood or where understanding is still limited can often be left out or understated

Tired about complaints about poor statistical analyses??

- Here’s a good one

Study charts 1,200-year history of Bear River

By Amy Joi O'Donoghue

February 18th, 2015 @ 7:01am

In canvas

ARTICLE IN PRESS

Journal of Hydrology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A millennium-length reconstruction of Bear River stream flow, Utah

R.J. DeRose [a,*], M.F. Bekker [b], S.-Y. Wang [c], B.M. Buckley [d], R.K. Kjelgren [c], T. Bardsley [e], T.M. Rittenour [f], E.B. Allen [g]

[a] USDA, Forest Service, Forest Inventory and Analysis, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, United States
[b] Department of Geography, 690 SWKT, Brigham Young University, Provo, UT 84602, United States
[c] Plant, Soil, and Climate Department, 4820 Old Main Hill, Utah State University, Logan, UT 84322-4820, United States
[d] Tree Ring Lab, Room 108, Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, NY 10964, United States
[e] Western Water Assessment, 2242 West North Temple, Salt Lake City, UT 84116, United States
[f] Department of Geology, 4505 Old Main Hill, Utah State University, Logan, UT 84322-4505, United States
[g] United States Geological Survey, 4200 New Haven Road, Columbia, MO 65201, United States

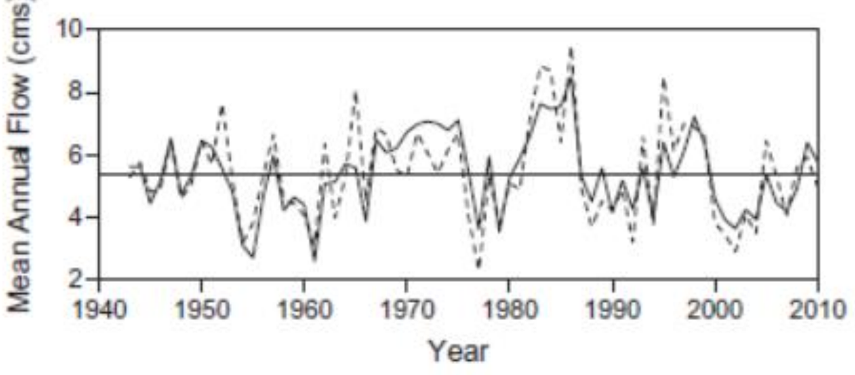


Fig. 2. Observed (dashed line) versus predicted (solid line) Bear River stream flow for the instrumental period (1943–2010). Horizontal line indicates instrumental mean water year flow (5.412 cms). Linear regression model explained 67% of the variation in instrumental Bear River flow.

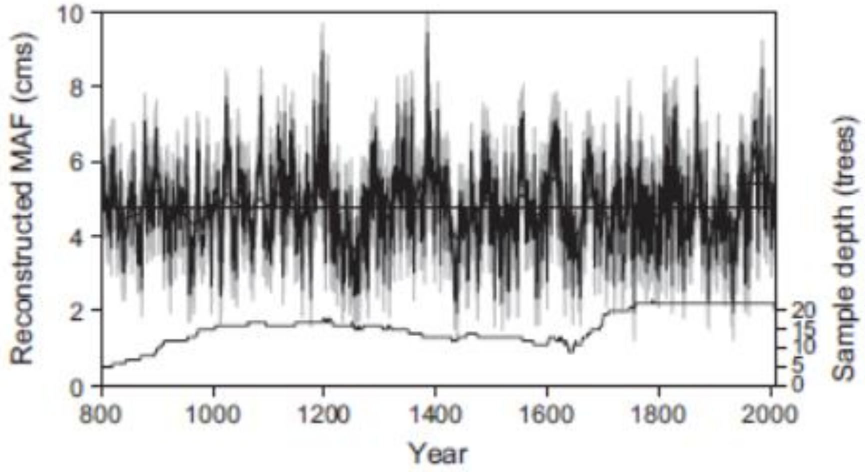


Fig. 3. Reconstructed Bear River stream flow from 800 to 2010 AD (thin black line), dark bold solid line cubic smoothing spline with 50% frequency cut-off at wavelength 20 years, light bold solid line cubic smoothing spline with 50% frequency at wavelength 60 years. Gray bands indicate 80% confidence interval calculated from the Bear River reconstruction model RMSE. Solid horizontal line is reconstructed MAF (4.796 cms). Dashed horizontal line is instrumental MAF (5.412 cms). Sample depth (number trees) for SFC indicated on the right.

Table 1
Model skill statistics and calibration–verification results for the Bear River reconstruction.

	r	R^2	Adj. R^2	RE	CE	Sign test (hit/miss)	RMSE (cms)
Calibrate (1943–1976)	0.72	0.52	0.50	0.66	0.39		
Calibrate (1977–2010)	0.90	0.81	0.80	0.23	0.13		
Full model	0.82	0.68	0.67			54/12[a]	0.8156

(r) – Pearson's correlation coefficient, (R^2) – coefficient of determination, (adj. R^2) coefficient of determination adjusted for degrees of freedom, RE – reduction of error statistic, CE – coefficient of efficiency statistic, RMSE – root mean-squared error.
Full model: 1.9414 + 2.9048 * SFC.
[a] Sign test significant at the alpha <0.01 level (Fritts, 1976).

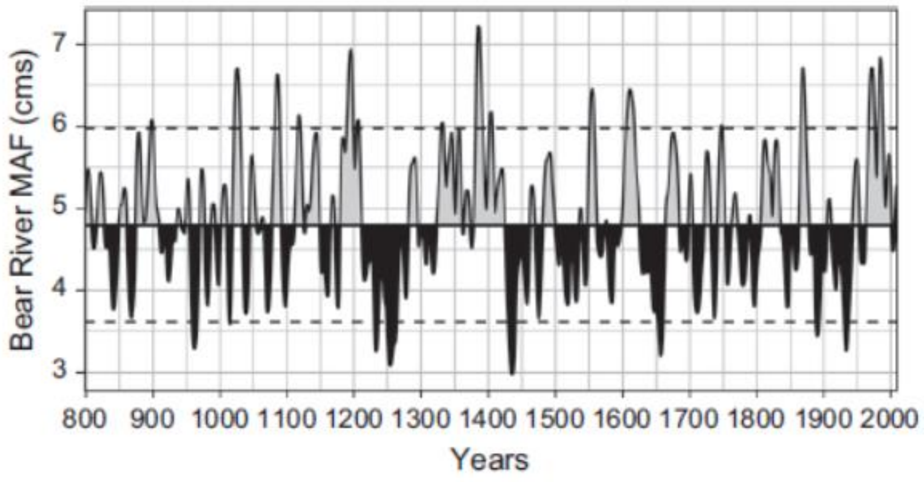
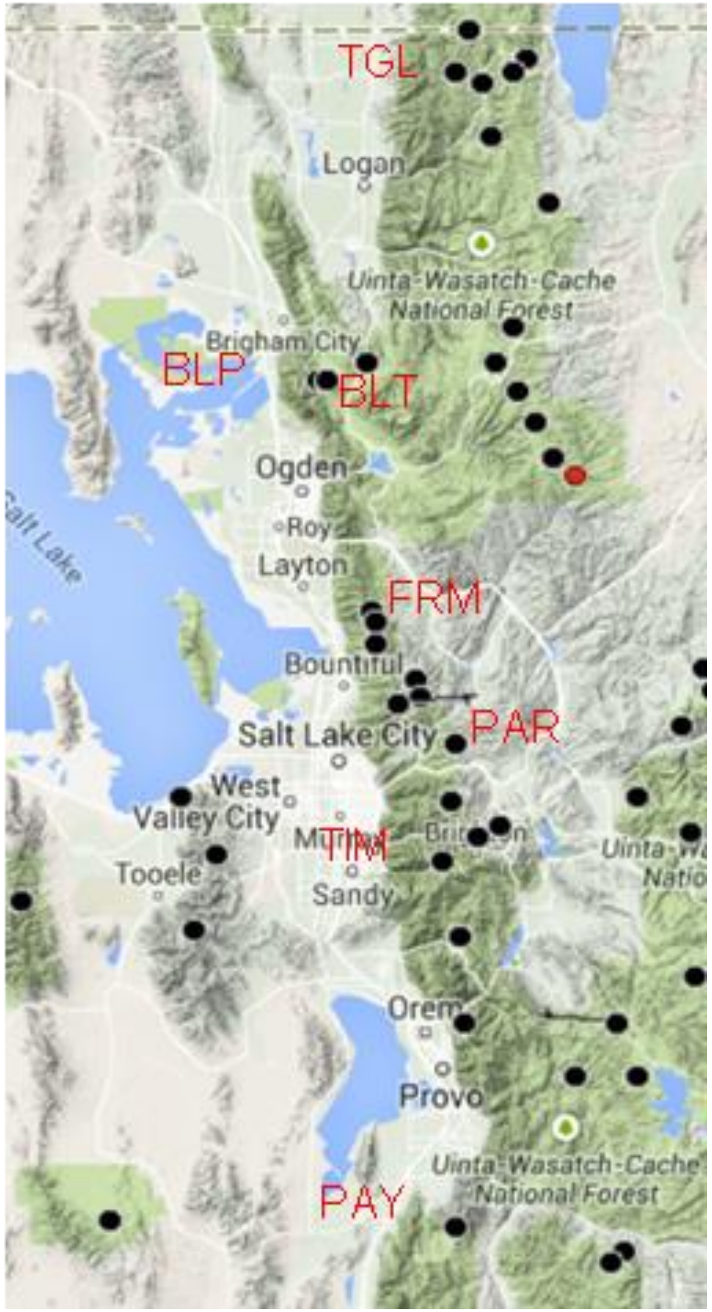


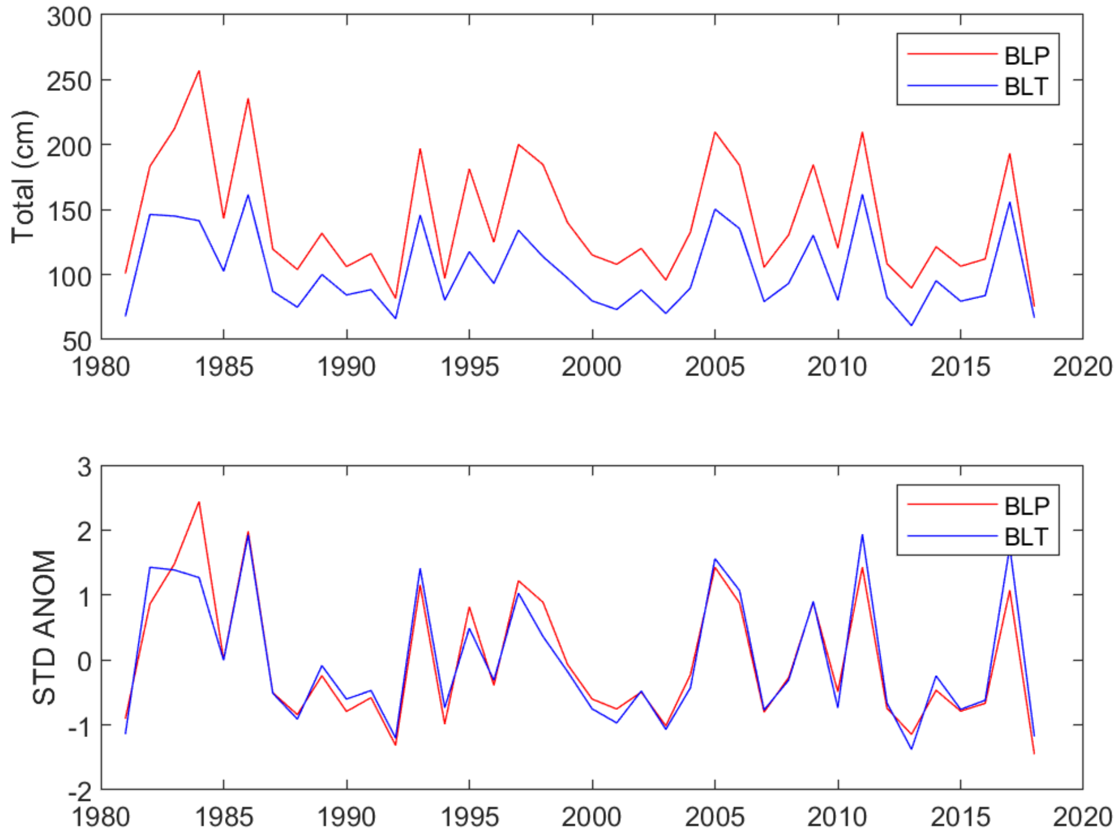
Fig. 5. Reconstructed Bear River decadal-scale drought (black) and pluvial (gray) periods from cubic smoothing spline with frequency response of 25% at wavelength 10 years. Dashed lines indicate 1 SD from reconstruction mean. See Table 3 for ranked dry and wet periods.

ior

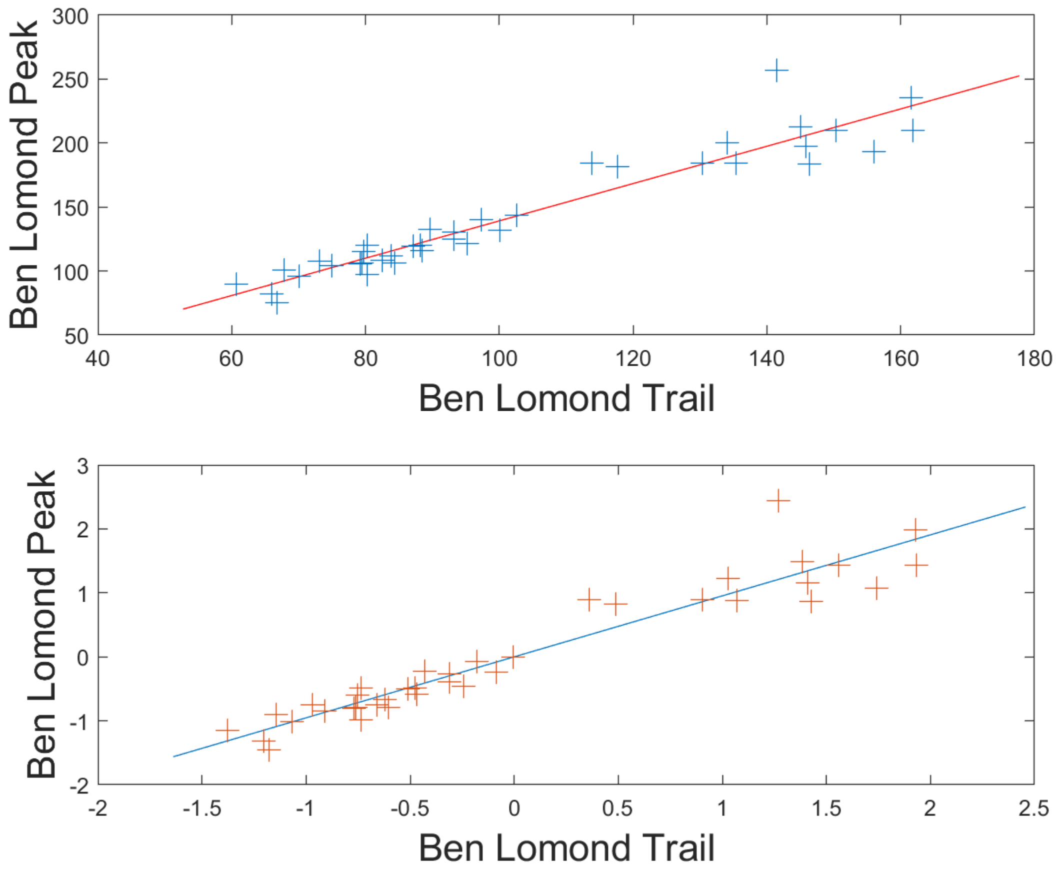
SNOTEL Sites



Accum Precip



Scatter plots



Estimating Values of One Variable From Another

- X- Ben Lomond Trail
- Y- Ben Lomond Peak
- Want to estimate Peak from Trail
- Use pairs of observations from sample
- Need to determine coefficient b or r
- b- slope of linear estimate
- r- linear correlation

$$\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$$

$$\hat{y}^*{}_i = r\hat{x}^*{}_i$$

Definitions

- Estimate $\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$
- Error of estimate $e_i = y_i' - \hat{y}_i'$
- Want $\sum_{i=1}^{n} e_i^2$ to be a minimum
- Need to find the value of b that minimizes that sum

$$\frac{\partial}{\partial b}\sum_{i=1}^{n} e_i^2 = 0$$

$$b = \overline{x_i' y_i'} / \overline{(x_i')^2} = \overline{x_i' y_i'} / s_x^2$$

- The value of b that minimizes the total error in the sample

Covariance

- Relates how departures of x and y from respective means are related
- Units are the product of the units of the two variables x and y
- Large and positive if sample tendency for:
  - large + anomalies of x occurring when large + anomalies of y AND
  - large - anomalies of x occurring when large - anomalies of y
- Large and negative if sample tendency for:
  - large + anomalies of x occurring when large - anomalies of y AND
  - large - anomalies of x occurring when large + anomalies of y
- Near zero when tendency for cancellation
  - large + anomalies of x occurring when both large – and + anomalies of y AND
  - large - anomalies of x occurring when both large – and + anomalies of y

Linear Correlation

$$r^2 = b^2 s_x^{\ 2} / s_y^{\ 2} = (\overline{x_i' y_i'})^2 / (s_x^{\ 2} s_y^{\ 2}) \qquad r = (\overline{x_i' y_i'}) / \sqrt{\overline{x_i'^2}\ \overline{y_i'^2}}$$

$$x_i^{\ *} = x_i' / s_x, y_i^{\ *} = y_i' / s_y, r = (\overline{x_i^* y_i^*})$$

$$1 = r^2 + \frac{\overline{e_i^2}}{s_y^{\ 2}}$$

y's total sample variance = fraction of variance estimated by x + fraction of variance NOT explained by x

- Dimensionless number relates how departures of x and y from respective means are related taking into account variance of x and y
- r = 1. Linear fits estimates ALL of the variability of the y anomalies and x and y vary identically
- r = -1 perfect linear estimation but when x is positive, y is negative and vice versa
- r = 0. linear fit explains none of the variability of the y anomalies in the sample. Best estimate of y is the mean value

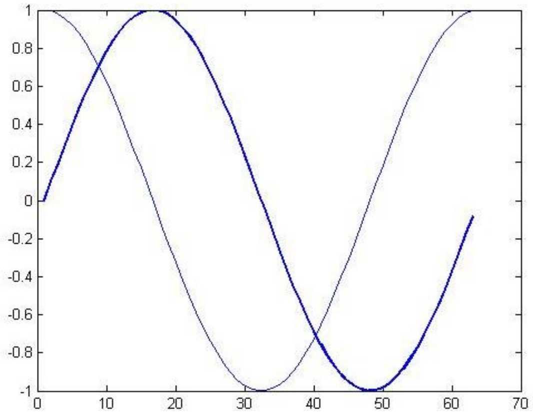
Linear Algebra is your friend

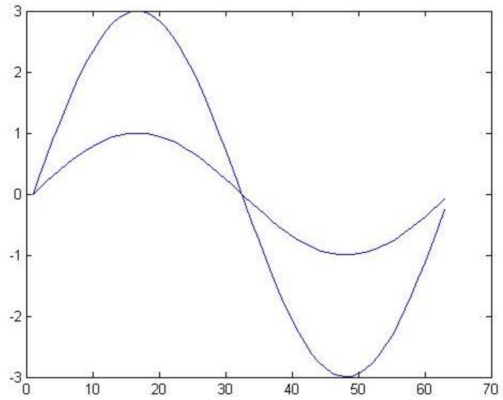
$$\vec{X}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \dots \\ x'_n \end{bmatrix} \quad \vec{Y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \dots \\ y'_n \end{bmatrix}$$

$$\overline{x'_i y'_i} = \vec{X}'^T \vec{Y}' / n$$

Stop and think before blindly computing correlations

- tendency to use correlation coefficients of 0.5 - 0.6 to indicate “useful” association.
  - 75%-64% of the total variance is NOT explained by a linear relationship if the correlation is in that range
- linear correlations can be made large by leaving in signals that may be irrelevant to the analysis. Annual and diurnal cyles may need to be removed
- large linear correlations may occur simply at random, especially if we try to correlate one variate with many, many others
- relationships in the data that are inherently nonlinear will not be handled well
- when two time series are in quadrature with one another then the linear correlation is 0
- Linear correlation provides no information on the relative amplitudes of two time series

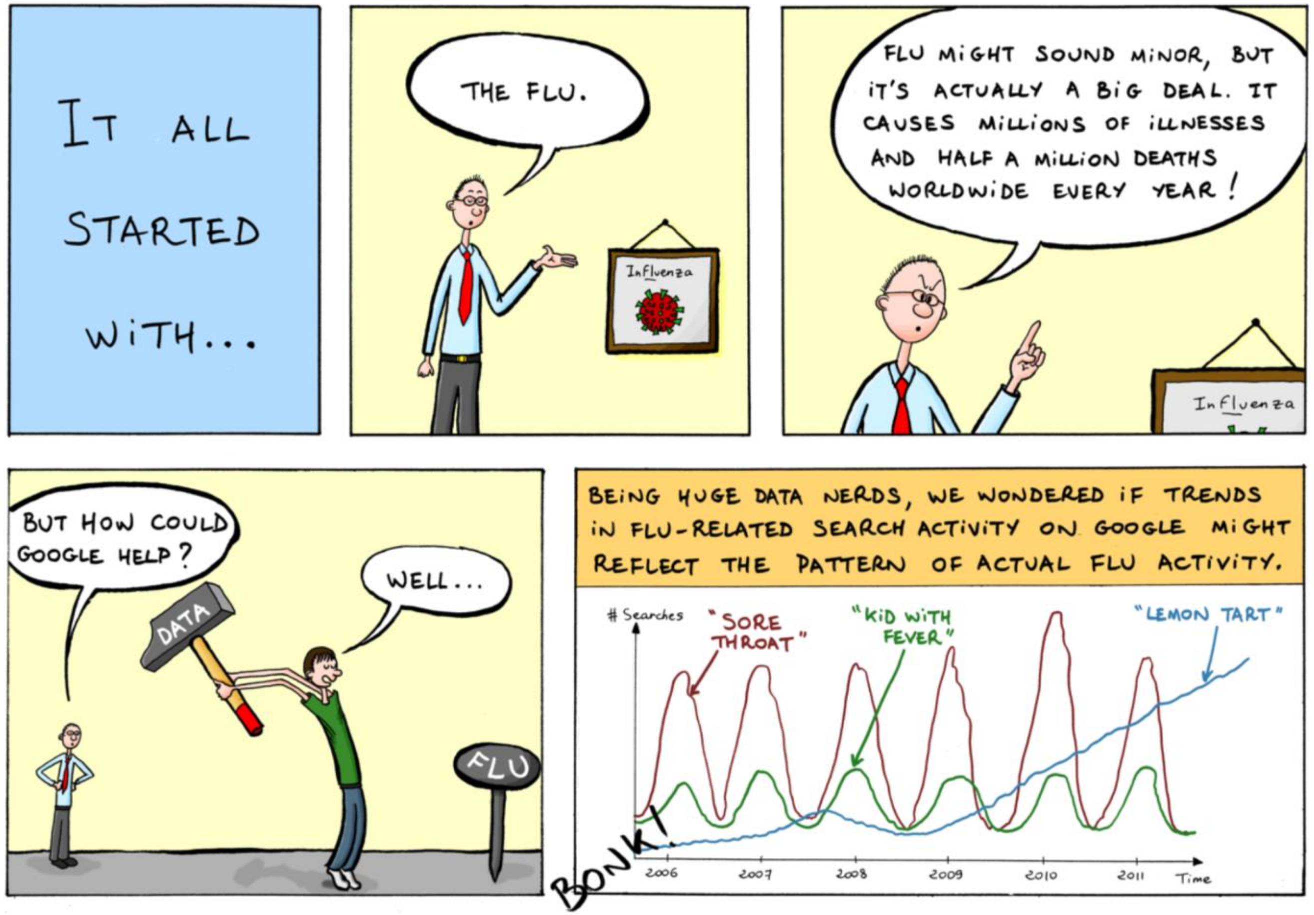




Getting things mucked up

- Incorrectly assuming causation from correlation
  - Having it backwards: wet ground causes rain
  - Something else is the underlying cause: decreased arctic ice will cause temperatures in Utah to increase (global warming?)

Google correlate



Resources for humor…

- https://www.google.com/trends/correlate
- https://www.google.com/trends/correlate/search?e=war&e=air+pollution&t=monthly&p=us#default,30

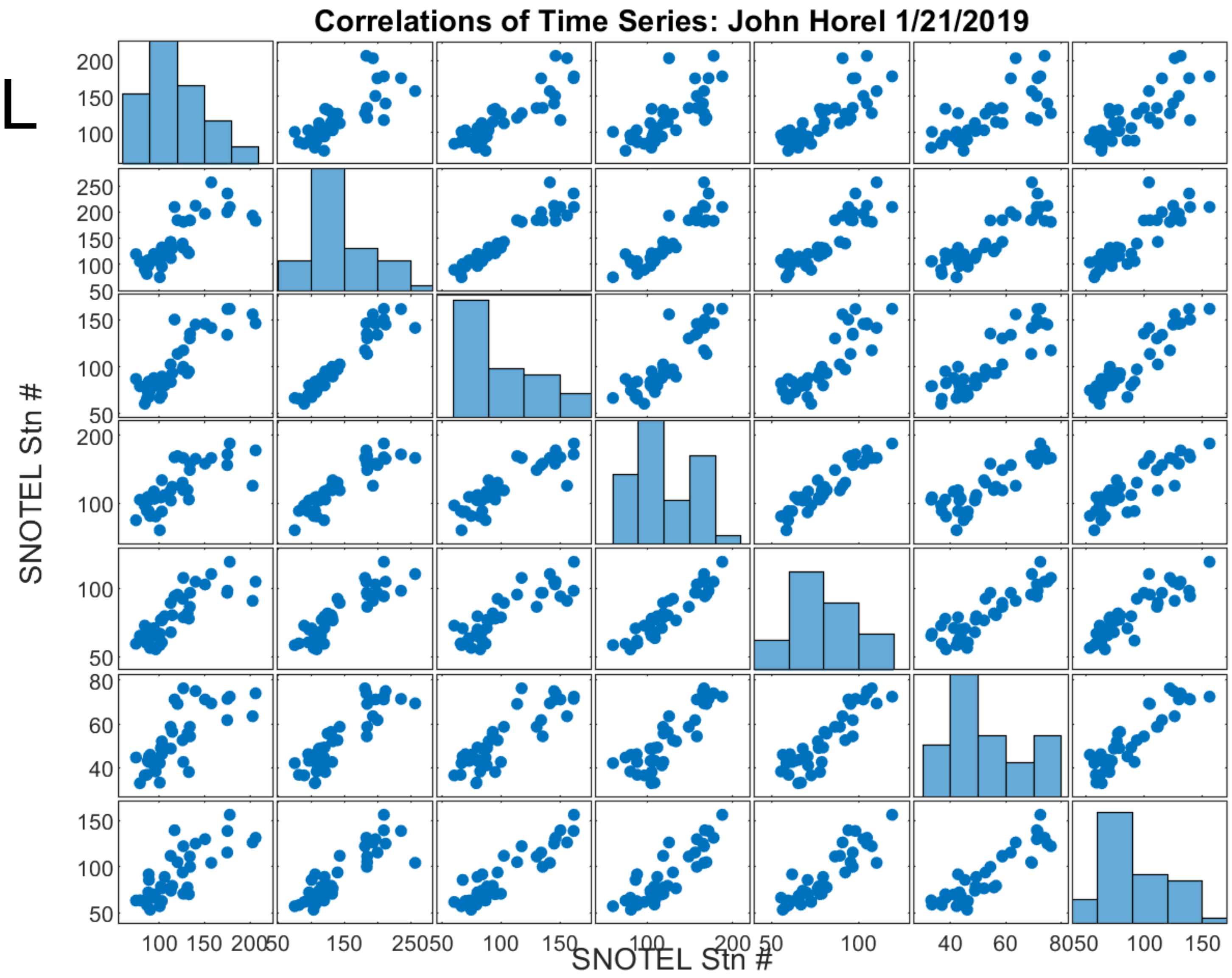
- http://tylervigen.com/spurious-correlations

Multivariate Linear Correlations

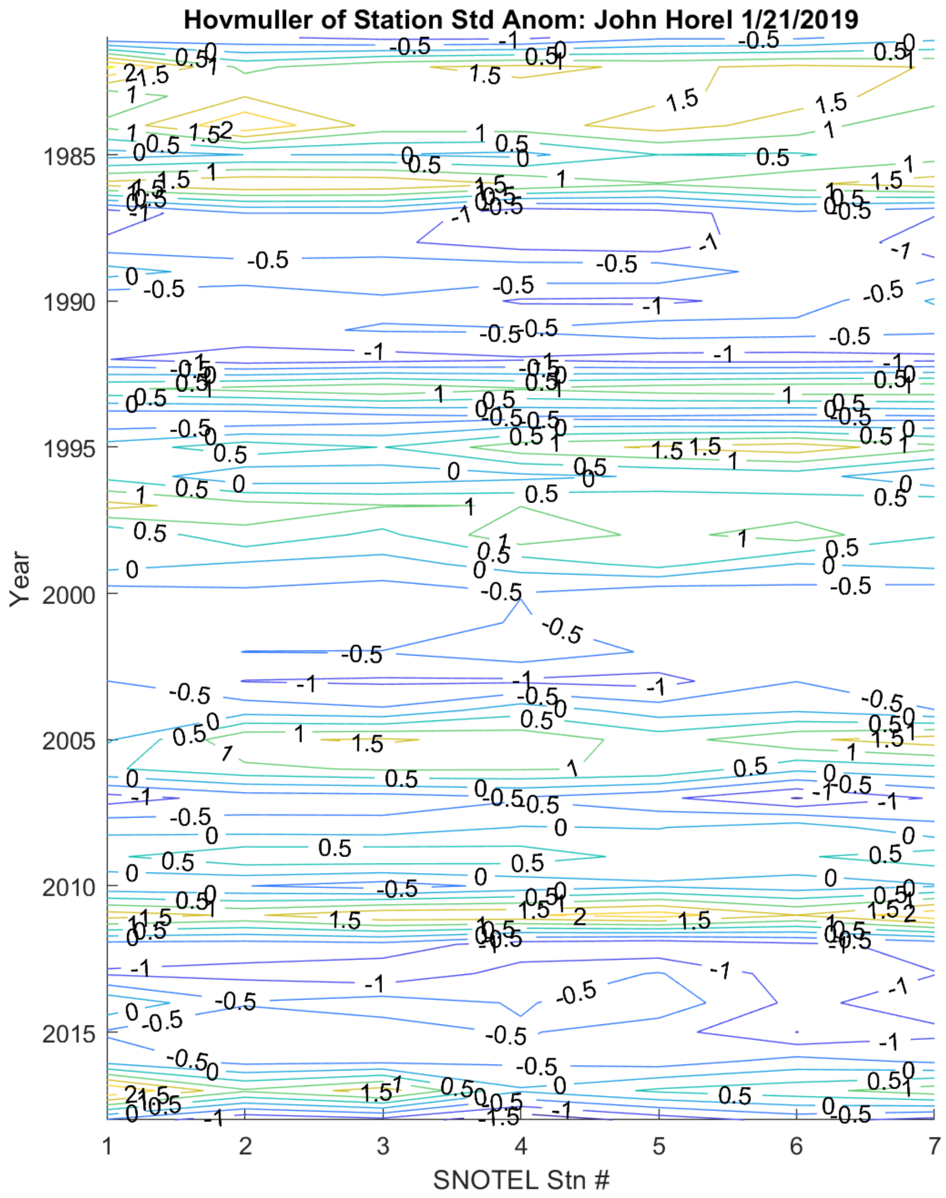
$$\vec{X}* = \begin{bmatrix} x*_{11} & x*_{12} & \dots & x*_{17} \\ x*_{21} & x*_{22} & \dots & x*_{27} \\ \dots & \dots & \dots & \dots \\ x*_{n1} & x*_{n2} & \dots & x*_{n7} \end{bmatrix}$$

- 7 stations and n=38 years
- Standardized anomalies

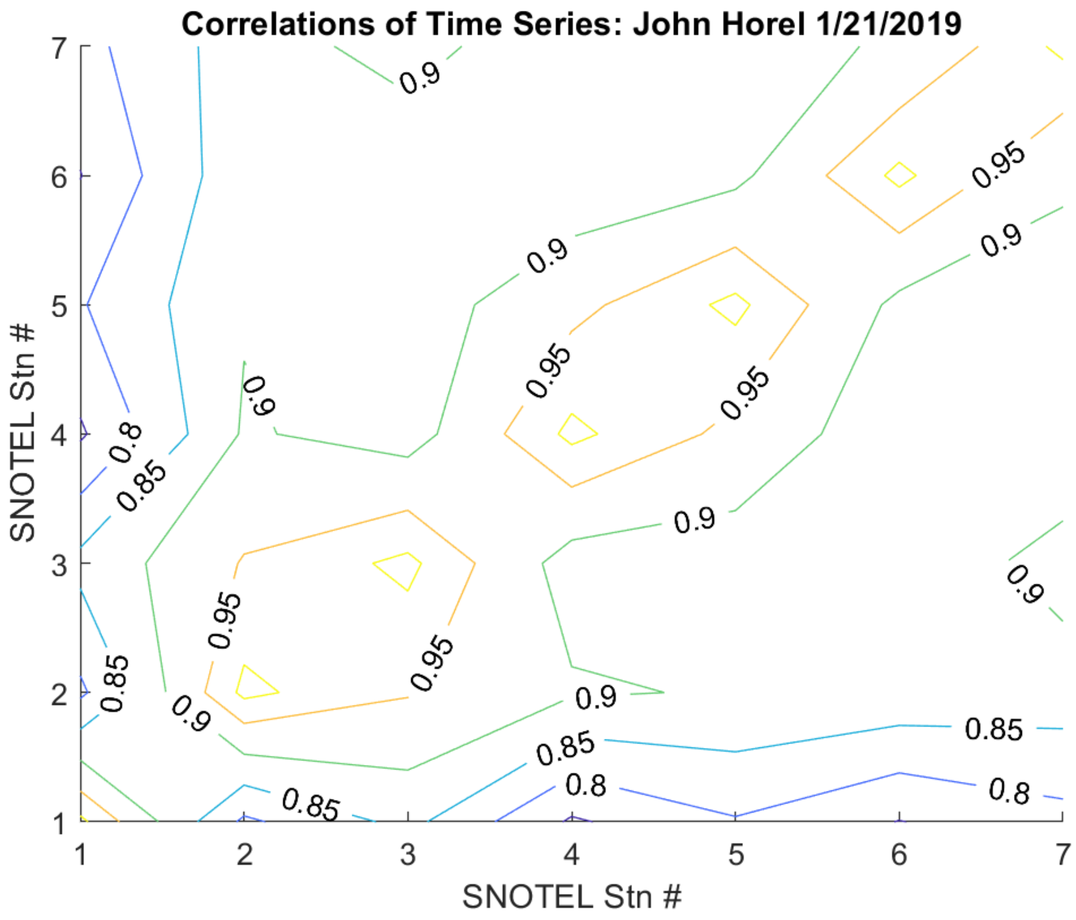
- TGL



TGL BLP BLT FRM PAR TIM PAY


Hovmuller of Station Std Anom: John Horel 1/21/2019
Year
SNOTEL Stn #
1985
1990
1995
2000
2005
2010
2015
1
2
3
4
5
6
7

Multivariate Linear Correlations



$$\vec{X}* = \begin{bmatrix} x*_{11} & x*_{12} & \dots & x*_{17} \\ x*_{21} & x*_{22} & \dots & x*_{27} \\ \dots & \dots & \dots & \dots \\ x*_{n1} & x*_{n2} & \dots & x*_{n7} \end{bmatrix}$$

$$\vec{R} = \vec{X}*^{T} \ \vec{X}*/n$$

Assignments

- Read Chapter 3 and 4 notes in Canvas (Chapter 5 also available)
- Read text: Text Chapter 3.1-3.6, Chapter 4
- Assignment 10 released soon