

FORTY YEARS OF WATER QUALITY STATISTICS: WHAT'S CHANGED, WHAT HASN'T?

DENNIS R. HELSEL

PRACTICALSTATS.COM



WHAT'S CHANGED?

STATISTICAL METHODS IN WATER RESOURCES 2nd Edition

- The sequel, with 3 new authors
- The 2nd edition will be published in 2019 [available in June]
- Will be a free download from:
USGS Publications: <https://pubs.er.usgs.gov>
Practical Stats:
<http://practicalstats.com/info2use/books.html>
- Discusses the five Changes in this talk, plus much, much more

STATISTICAL METHODS IN
WATER RESOURCES
2ND EDITION

Helsel, Hirsch, Archfield,
Ryberg & Gilroy

USGS Techniques and Methods
4-A3
(2019)



WHAT HASN'T CHANGED #1

“IT’S ROBUST”

USE t-TEST & ANOVA, IT’S ROBUST

- Montgomery and Loftis (1987)
- Johnson (1995)
- Knief and Forstmeier (Dec 2018)

PROBLEMS

- Lack of Power -- “Robust” considers only false positives, not false negatives
- Transformations Change What Is Being Tested
- Cannot Use Censored Data (Nondetects)
- Mean Doesn’t Represent the Center of Skewed Data Very Well
- Reliance on the Central Limit Theorem for $n < 70$ (or 100?) Is Unwise



WHAT'S CHANGED #1

Permutation Tests are More Powerful

- Permutation Tests: testing difference in means without assuming normality. Distribution-free.
- Bootstrapping: computing confidence intervals on the mean without assuming normality (without t coefficients)
- Hahn and Meeker (1991): “One might ask ‘When should I use distribution-free statistical methods?’ The answer, we assert, is ‘Whenever possible.’ If one can do a study with minimal assumptions, then the resulting conclusions are based on a more solid foundation.”
- Example: t-test of difference in means of 2 groups (n=16): p-value

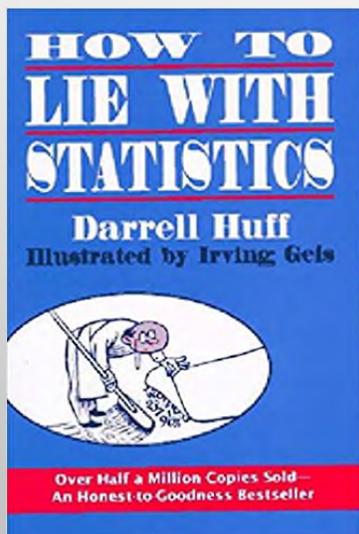
t-test	0.14 (doesn't find differences that are there)
Permutation test on means:	0.0018
Wilcoxon rank-sum test on percentiles:	0.01



WHAT HASN'T CHANGED #2

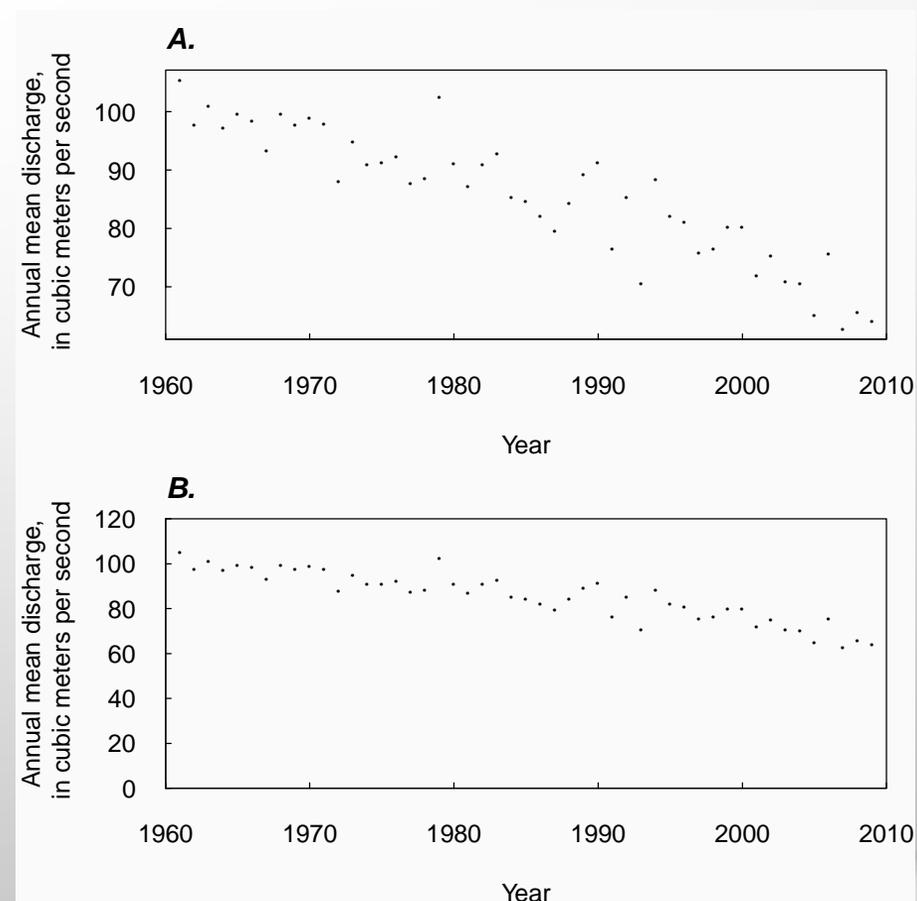
PLOTS THAT INCORRECTLY MAXIMIZE IMPORTANCE

HUFF (1953)



If you start your plot at the minimum of your data, any changes will look large. -- Huff

Solution:
Axes should start at zero when absolute magnitude is important (Helsel, Hirsch, Archfield, Ryberg and Gilroy (2019))



WHAT HASN'T CHANGED #3

DELETING OUTLIERS FOR NO REASON

OUTLIER DELETION WITH NO JUSTIFICATION

- “Excluding the outlier samples, the annual average detected concentration of MTBE ranges from.....” -- circa 2008
Consultant's report
- “This city in Alaska is warming so fast, algorithms removed the data because it seemed unreal” -- Denver Post, 12/12/17
Computer algorithm
- a) Delete any observations designated as outliers by Rosner's test -- 2014 USEPA guidance; b) 2011 USEPA report removed outliers failing the outlier test after transforming data to make data LESS normal.
Government report/memo

PROBLEMS

- There is no test for “bad data” in statistics
- Outlier tests determine if observations likely came from a normal distribution -- water, air, soils and chemical data rarely do
- If an outlier is negatively affecting your statistic or test, you are probably using the wrong statistic/test



WHAT'S CHANGED #3

THE ISSUE IS GETTING MORE ATTENTION

- Outlier deletion has become a somewhat frequent topic in court cases. Is there scientific reason for deletion? Basing the decision on the dataset itself is not sufficient reason. Deleting outliers (such as high concentrations) may miss important conditions (contamination, high flows). The company/org/person may have to explain why they deleted them. What do statisticians and leading scientists think?
- Barry Nussbaum, formerly Chief Statistician of USEPA: “There are a lot of statistical methods looking at whether an outlier should be deleted I don’t endorse any of them.”
- Ed Gilroy, formerly Statistician at USGS: “Treat outliers like children correct them when necessary, but never throw them out.”
- Marcia McNutt, Editor-in-Chief of Science: “Clearly, throwing out a few of the data points by declaring them ‘outliers’ would have improved the fit dramatically....It was not too long before it was realized that those ‘outliers’ were the key to a more complete understanding of the long-term rheological behavior of the oceanic plates.” (Raising the Bar, 2014 Editorial).



WHAT HASN'T CHANGED #4

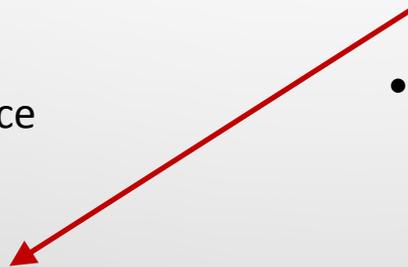
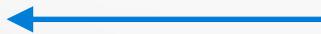
OVER-RELIANCE ON p-VALUES; p-HACKING

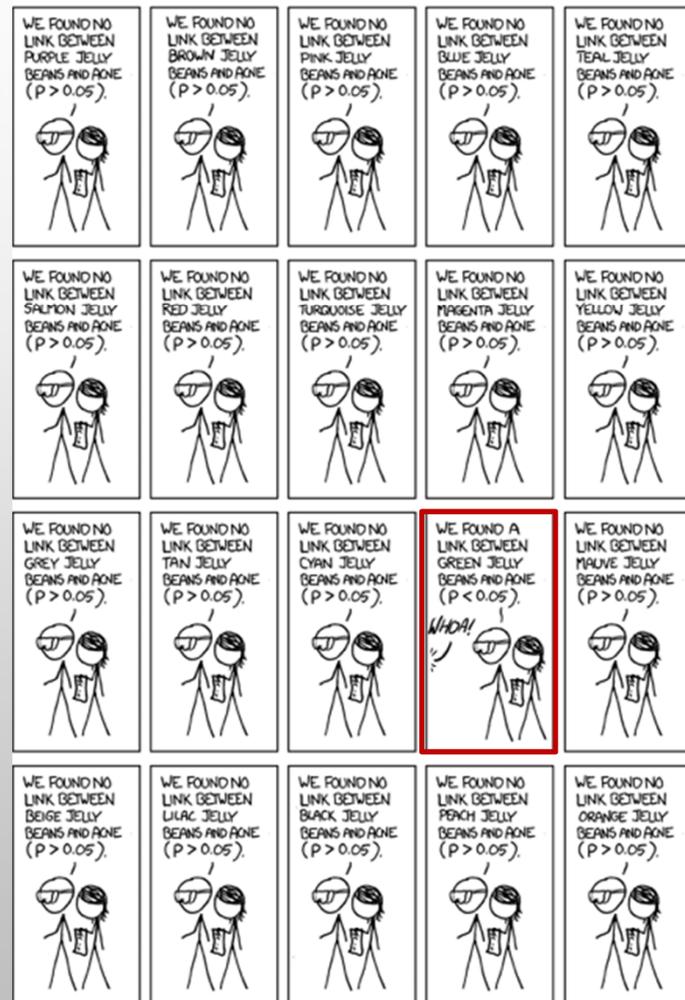
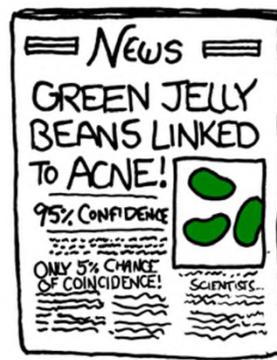
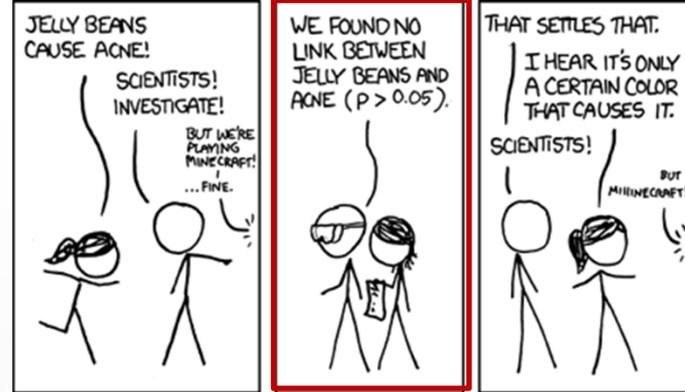
MISCONCEPTIONS

- "A large p-value proves the null hypothesis"
 - may be due to few data or tests with low power.
 - absence of evidence is not evidence of absence.
- "A significant p-value indicates practical usefulness"
 - the effect may be small enough to have no human or ecological effects.

ISSUES

- p-values are a function of sample size.
 - few data: large trends may not be seen
 - lots of data: unimportant trends found
- Researchers sometimes try multiple hypothesis tests, removal of outliers, deletion of groups, etc. to achieve statistically significant results. This process is termed "p-hacking".
- Some journals balk at publishing results with $p > 0.05$. No change is sometimes the most welcome result or can inform decision makers that some action did not have the hoped-for results.





p-Hacking

an α of 0.05 means that there is a 1 in 20 probability of a false positive. Don't just keep trying until you get a significant result!

Figure from xkcd.com (Munroe, 2016), used under creative commons attribution-noncommercial license



WHAT IS CHANGING #4a

FLEXIBLE TREND ANALYSIS

WRTDS smoothing (Hirsch et al., 2010. JAWRA 46:5, 857-880)

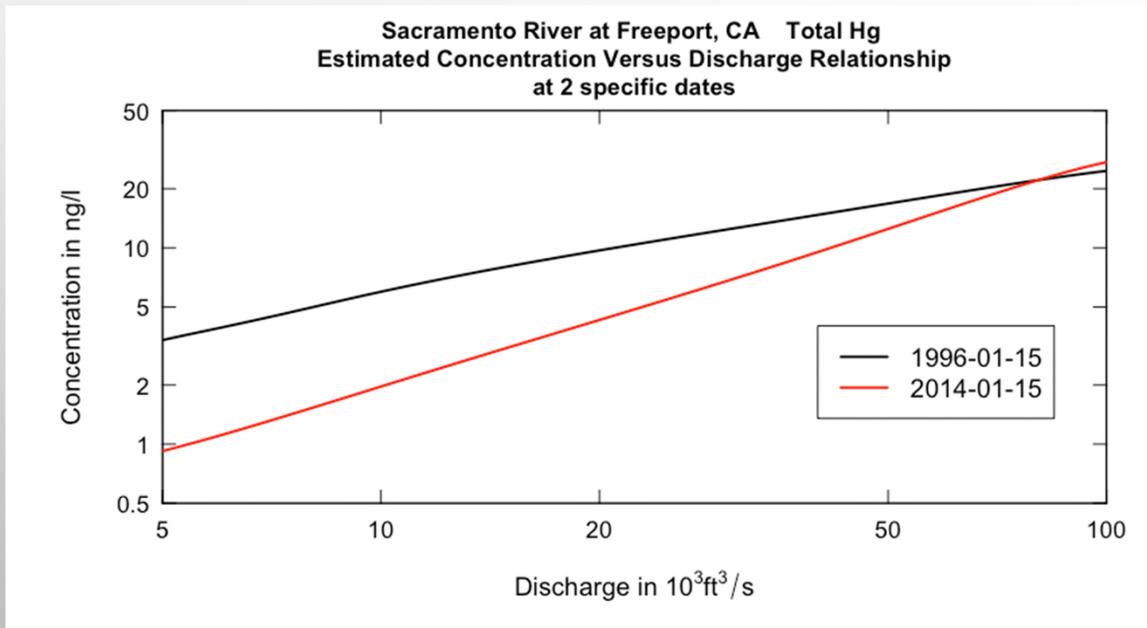
- Exploratory and quantitative
- Concentration vs Flow relationship can change over time
- Seasonal pattern can change over time
- Temporal pattern of any shape, including non-monotonic
- Analysis of both concentration trends & flux trends
- Makes estimates of actual history & flow-normalized history
- Handles censored data, “less-thans”
- Handles non-stationary discharge history
- Has been used to estimate frequency of exceedances
- Reports the uncertainty of trend estimates



WHAT IS CHANGING #4a

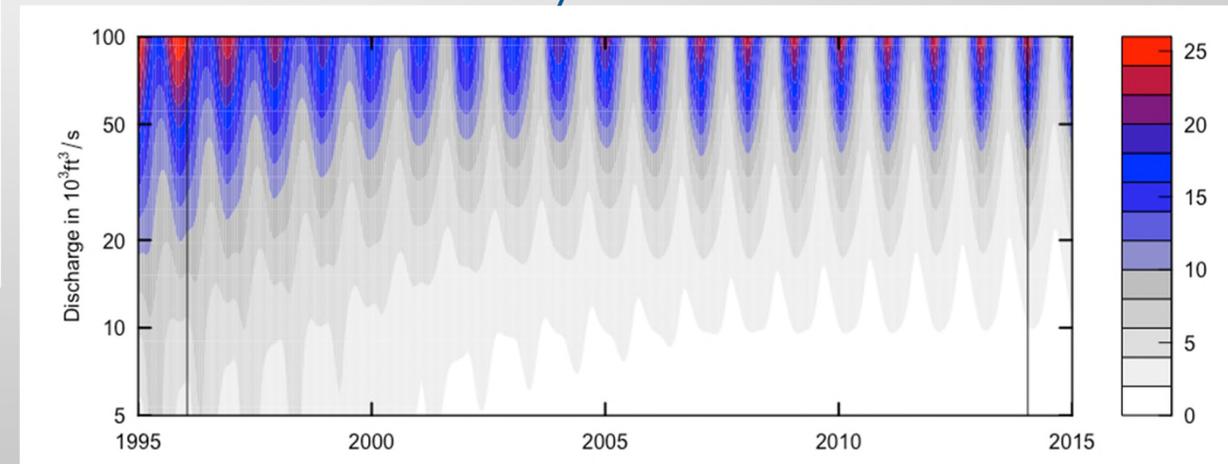
FLEXIBLE TREND ANALYSIS

Standard regression would force these two lines to be parallel.



Total Mercury Decreases 1995-2015. WRTDS uses statistical smoothing to estimate $E[\text{Conc}] = f(\text{Discharge})$ for any given date. Fourfold decrease at lower discharges, none at higher.

Contour plot shows the $E[\text{Conc}] = f(\text{Discharge})$ for each of the 7300 days in this record



WHAT IS CHANGING #4b

FLEXIBLE TREND ANALYSIS

Trend Detection Assessment (TDA)

- McBride, G. et al., Environ Monit Assess 186:5, 2729–2740 (2014); McBride, G. Journ. of Env. Quality 48: 2, 412-420 (2019)
- Focus on the direction of trend and how certain that direction is. Determination of magnitude is a function of whether there is sufficient data. With few data, large trends may go unnoticed. Instead, compute the probability that the slope is not zero.
- Use Bayesian Credible Intervals on the slope. Report with a graduated scale: “very likely” if 90-99% intervals do not include zero. Communicate the certainty that the slope is increasing (or decreasing if negative). If zero is included in the interval, report as “direction cannot be determined”.
- Categories with lower certainty would be “somewhat likely”. These statements give decision-makers more guidance than just ‘a trend is significant at $p < 0.05$ ’. But does not provide an estimate of magnitude for the slope. McBride and others (2014) state that decision-makers primarily want to know ‘are things improving?’, ‘is quality being maintained?’.



WHAT HASN'T CHANGED #5

USING EXCEL FOR STATISTICAL COMPUTATIONS

WHAT TYPES OF ANALYSIS CAN EXCEL NOT DO?

- Spearman's and Kendall's rank correlation coefficients
- 2-way ANOVA with unequal sample sizes (unbalanced data)
- Multiple comparison tests (post-hoc tests following ANOVA)
- Levene's test for equal variance (the older F-test used in Excel is far less accurate)
- Nonparametric tests, including the rank-sum, Kruskal-Wallis and Friedman tests
- Regression diagnostics, such as Mallow's Cp and PRESS (Excel does compute adjusted r-squared and standardized residuals)
- Survival analysis methods (for nondetects)
- Tests for serial correlation
- LOESS smooths

WHAT DOES EXCEL DO INCORRECTLY?

- Regression residuals Normal Probability Plot option. Draws a uniform distribution probability plot, even though it is labelled as a Normal Probability Plot. The plot is therefore useless and misleading for judging the adequacy of regression residuals.
- Excel's regression residuals plots use the original data rather than predicted values on the X axis. This is acceptable for simple regression with one X variable, but not for multiple regression.

Excel does not include modern methods for statistical analysis



WHAT'S CHANGED #5

SOFTWARE FOR MODERN STATISTICS

- PAST (free)
Performs nonparametric and permutation tests, regression diagnostics, some multivariate methods. Pull-down menus and easily learned.
- Commercial Software
incorporating newer methods such as bootstrapping and permutation tests. Easier to use than R for part-time data analysts. Residuals analysis for regression is excellent.
- R (free)
World's standard in statistics. Performs anything you can think of and more. Newly developed methods are more often found here than anywhere else. Does have a learning curve similar in difficulty to SAS.



STATA



...

R

Statistical Methods in Water Resources, 2nd edition uses R for all examples. Scripts for computations and all code for all figures may be downloaded.



SUMMARY

1. Use permutation tests and bootstrapping. You will miss signals if you continue to use old parametric tests and confidence intervals.
2. Take the y-axis down to zero whenever the numeric scale is meaningful.
3. Do not delete outliers unless you have evidence outside of the dataset showing they are in error or from another population.
4. Don't "p-hack" or overly rely on p-values. Understand your data using graphical procedures such as WRTDS to comprehend the full story.
5. Use modern statistical software. If R seems too complex, try PAST.
6. Download the second edition of *Statistical Methods in Water Resources* when it becomes available in June !!



THANK YOU FOR ATTENDING

- Most material is based on the book by Helsel, Hirsch, Archfield, Ryberg and Gilroy (2019).
- All opinions are my own and do not represent those of anyone else you can think of.

- Questions?

Get in touch!

Dennis Helsel dhelsel@practicalstats.com

Courses & free webinars at: <http://practicalstats.com/training>
(next webinar April 23rd)

