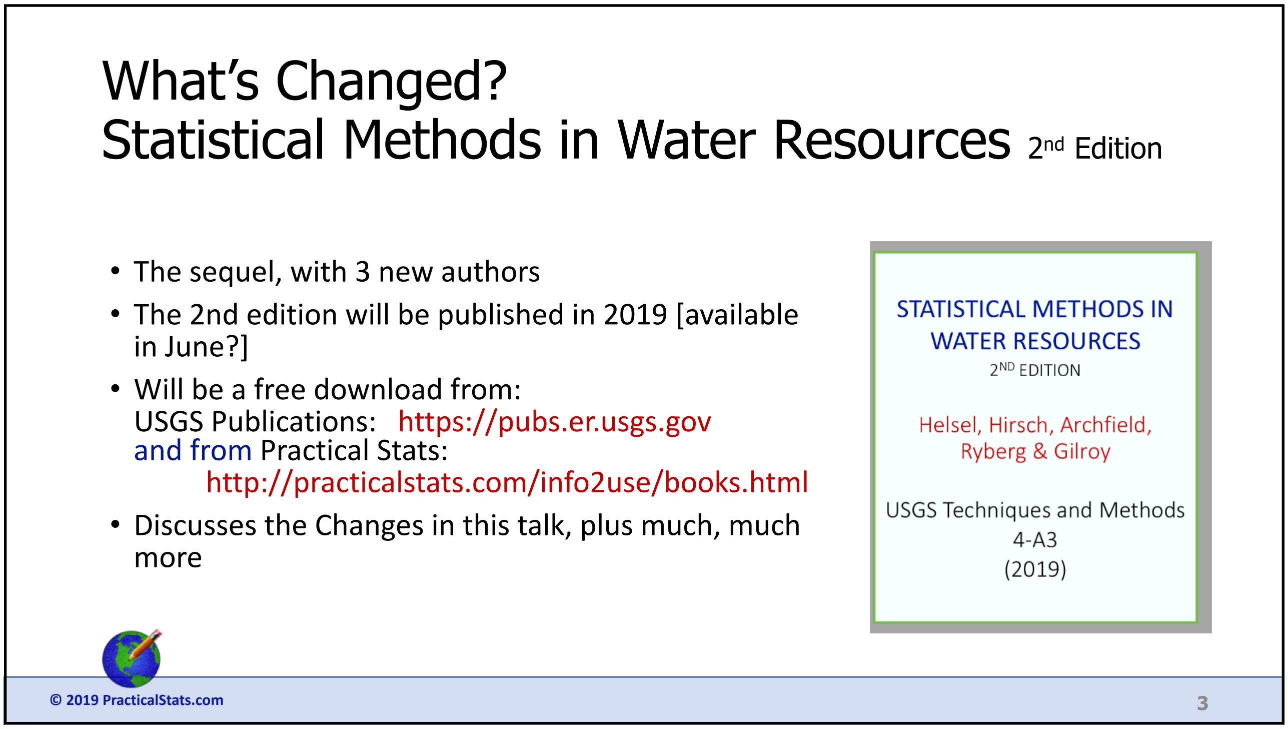
Forty Years of Water Quality Statistics: What's Changed, What Hasn't?

Dennis R. Helsel

PracticalStats.com

Objectives of the ‘Forty Years’ webinar

1. To present 5 bad habits / misunderstood practices that after 40 years are still commonly done today (the 5 “zombies”) in environmental sciences
2. To present current methods that have replaced the zombies in statistical practice and should quickly replace them in your work
3. To discuss one or two directions that statistical practice may be heading in the future, and if so will impact your work



What's Changed? Statistical Methods in Water Resources 2nd Edition

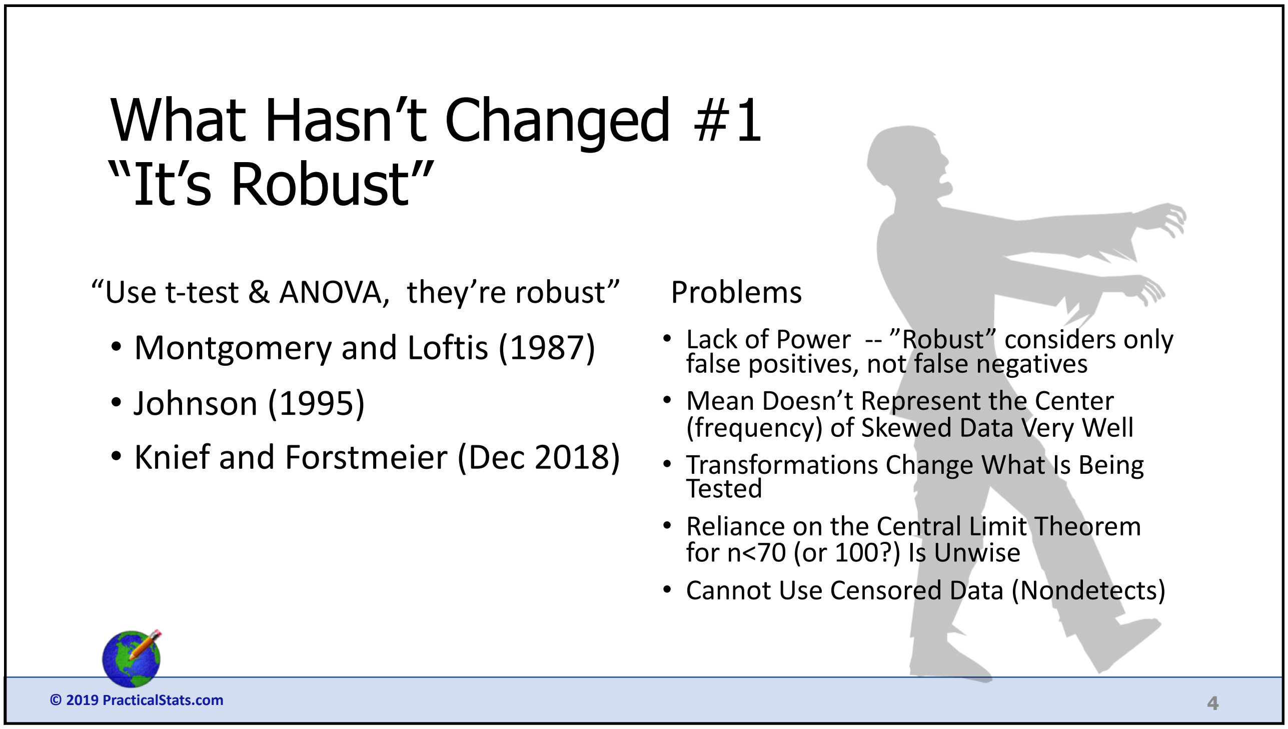
- The sequel, with 3 new authors
- The 2nd edition will be published in 2019 [available in June?]
- Will be a free download from:
  USGS Publications: https://pubs.er.usgs.gov
  and from Practical Stats:
  http://practicalstats.com/info2use/books.html
- Discusses the Changes in this talk, plus much,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

STATISTICAL METHODS IN WATER RESOURCES
2ND EDITION

Helsel, Hirsch, Archfield, Ryberg & Gilroy

USGS Techniques and Methods
4-A3
(2019)

3



What Hasn't Changed #1 "It's Robust"

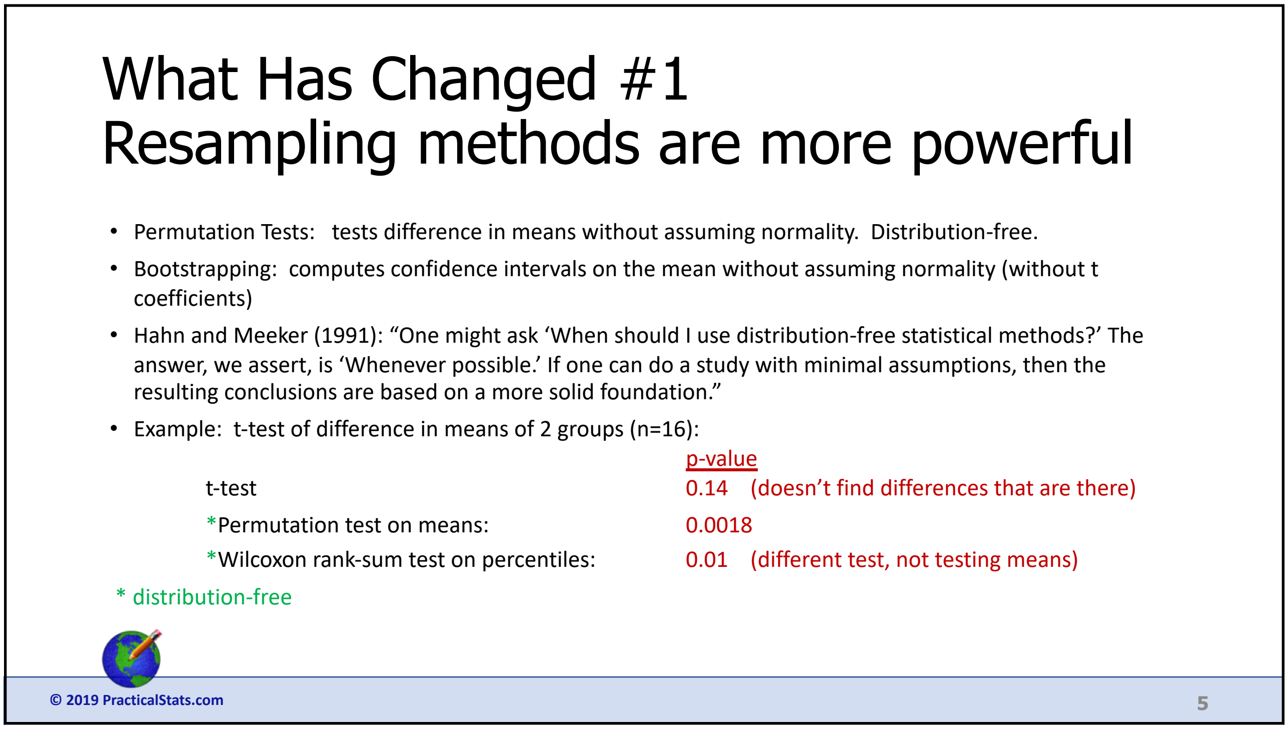
"Use t-test & ANOVA, they're robust"

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

Problems

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

4



What Has Changed #1
Resampling methods are more powerful

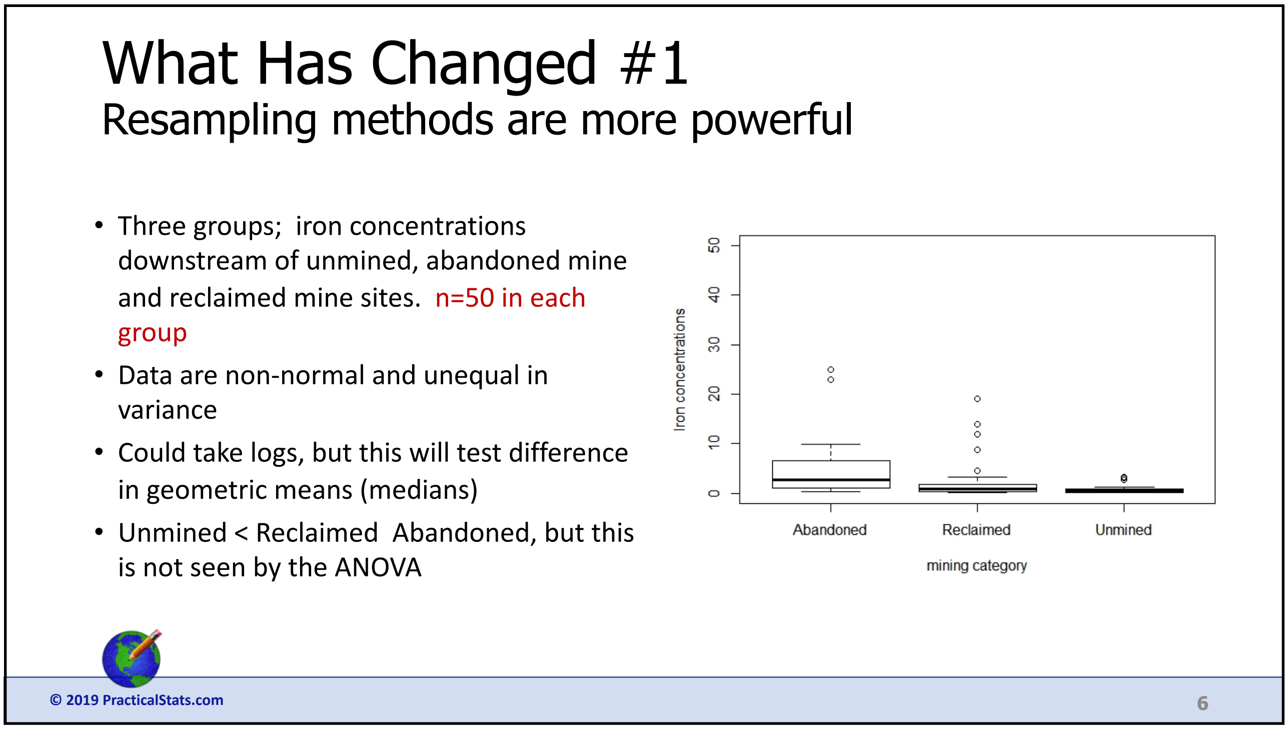
- Permutation Tests: tests difference in means without assuming normality. Distribution-free.
- Bootstrapping: computes confidence intervals on the mean without assuming normality (without t coefficients)
- Hahn and Meeker (1991): “One might ask ‘When should I use distribution-free statistical methods?’ 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 (different test, not testing means)

* distribution-free

© 2019 PracticalStats.com

5



What Has Changed #1
Resampling methods are more powerful

- Three groups; iron concentrations downstream of unmined, abandoned mine and reclaimed mine sites. n=50 in each group
- Data are non-normal and unequal in variance
- Could take logs, but this will test difference in geometric means (medians)
- Unmined < Reclaimed Abandoned, but this is not seen by the ANOVA

© 2019 PracticalStats.com

6

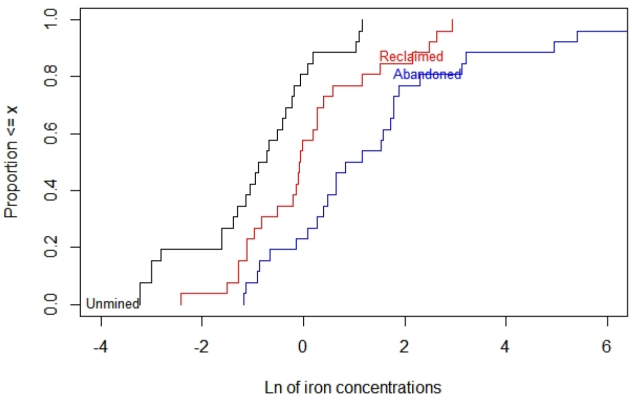
What Has Changed #1
Resampling methods are more powerful

ANOVA p = 0.10

Permtest p = 0.001

Same objective (is there a difference in means?), same data.

The difference is an indication of the loss of power of the parametric test. Sample size of 50 in each group is insufficient to overcome the power loss.



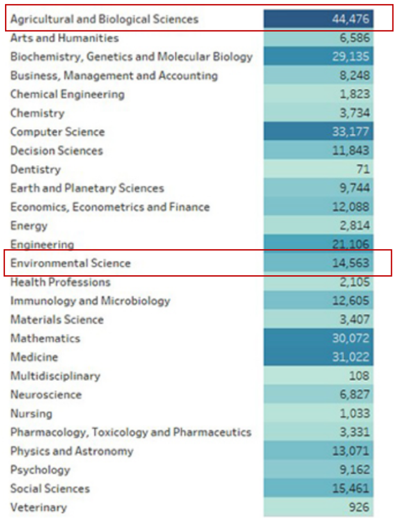
What's Changed #1
Resampling methods are more powerful

The 2018 International Prize in Statistics was awarded to Bradley Efron, professor of statistics and biomedical data science at Stanford University, in recognition of the "bootstrap," a method he developed in 1977 for assessing the uncertainty of scientific results that has had extraordinary impact across many scientific fields.

Other people are using it. Are you?

The impact of the bootstrap across research fields as measured by citation

The dataset contains over 200,000 articles from over 200 journals between 1980 and 2018



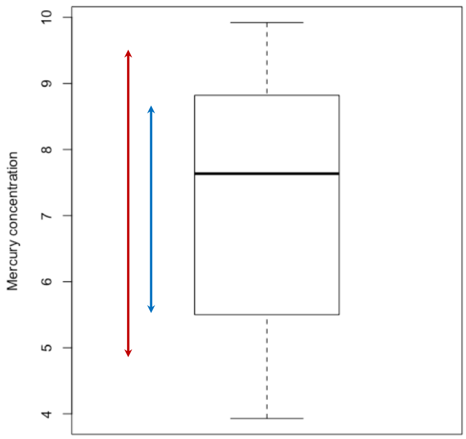
Agricultural and Biological Sciences	44,476
Arts and Humanities	6,586
Biochemistry, Genetics and Molecular Biology	29,135
Business, Management and Accounting	8,248
Chemical Engineering	1,823
Chemistry	3,734
Computer Science	33,177
Decision Sciences	11,843
Dentistry	71
Earth and Planetary Sciences	9,744
Economics, Econometrics and Finance	12,088
Energy	2,814
Engineering	21,106
Environmental Science	14,563
Health Professions	2,105
Immunology and Microbiology	12,605
Materials Science	3,407
Mathematics	30,072
Medicine	31,022
Multidisciplinary	108
Neuroscience	6,827
Nursing	1,033
Pharmacology, Toxicology and Pharmaceutics	3,331
Physics and Astronomy	13,071
Psychology	9,162
Social Sciences	15,461
Veterinary	926

What Has Changed #1
Resampling methods are more powerful

- Bootstrapped confidence intervals avoid the t-interval assumptions of normality and therefore symmetry.
- Lower end is often shorter with bootstrap because a t-interval uses std dev inflated by high outliers for the low end as well
- Example: Mercury concentrations. Compute the 95% CI

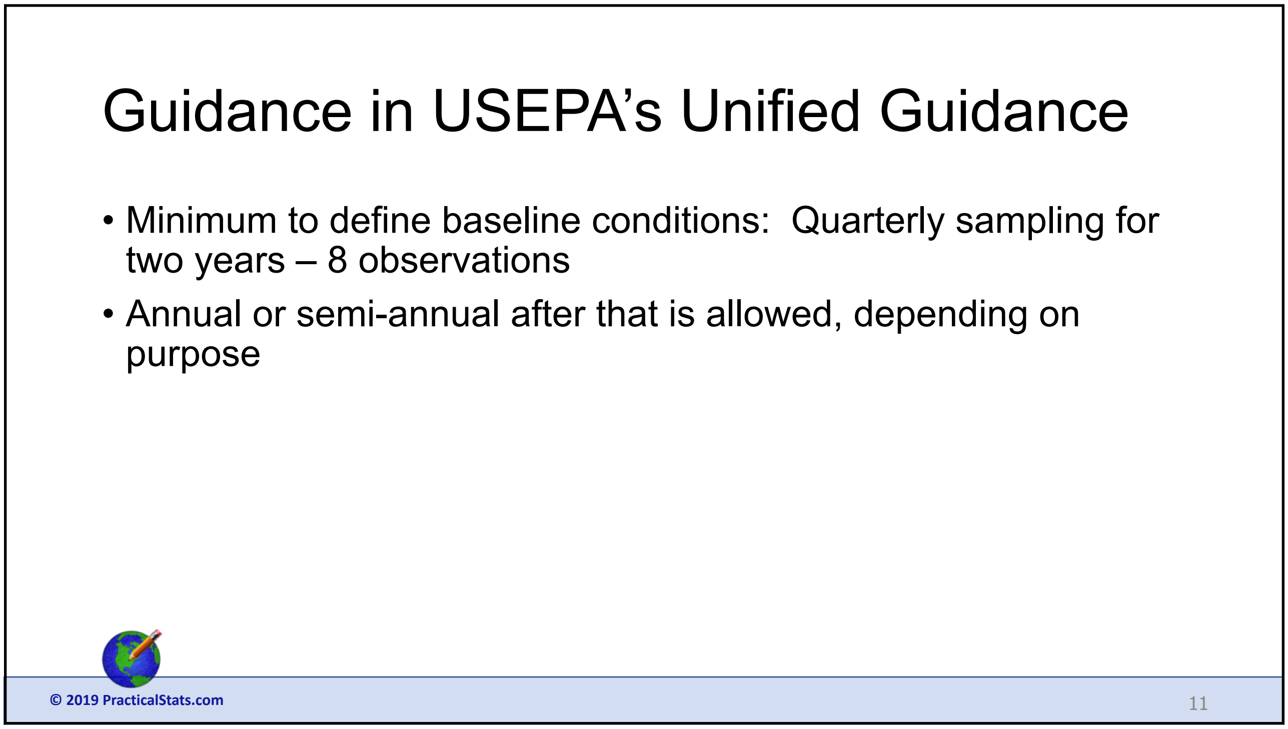
t interval: 4.94 to 9.54

bootstrap: 5.59 to 8.79 (shorter)



What Hasn't Changed #2
"How few observations can I get away with?"

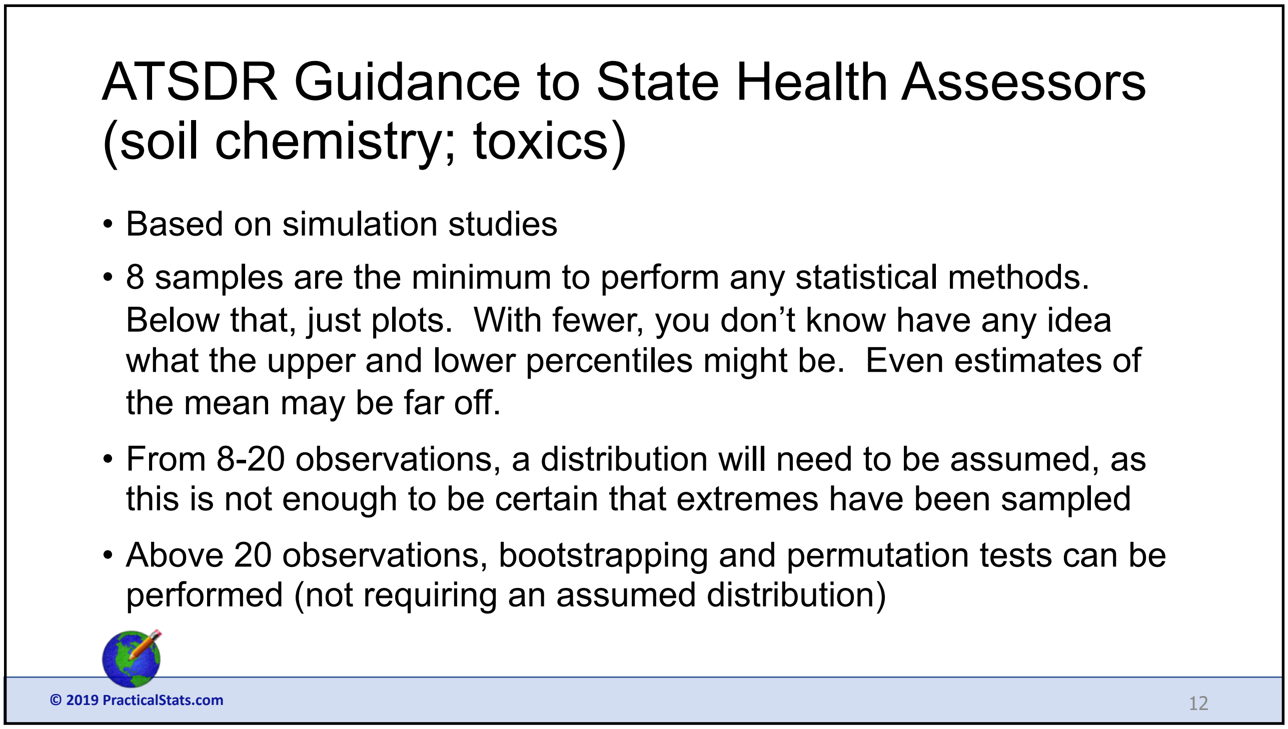
- Older guidance documents sometimes refer to "large" sample sizes as n≧30 (such as EPA QA/G9s, 2008). This is not "large"! This 'urban legend' invites people to collect fewer samples
- Groundwater quality studies are often done with 2 samples per year. After the 2nd year, trends are to be computed! For surface water quality, trends are computed after 5 years of annual sampling (are many issues with that, but sample size is one of them)
- Frequent guidance says to compute a UCL95 for small datasets, and if the computed value exceeds the current maximum, use the current maximum as the UCL. This likely is a strong underestimate of the UCL.



Guidance in USEPA’s Unified Guidance

- Minimum to define baseline conditions: Quarterly sampling for two years – 8 observations
- Annual or semi-annual after that is allowed, depending on purpose

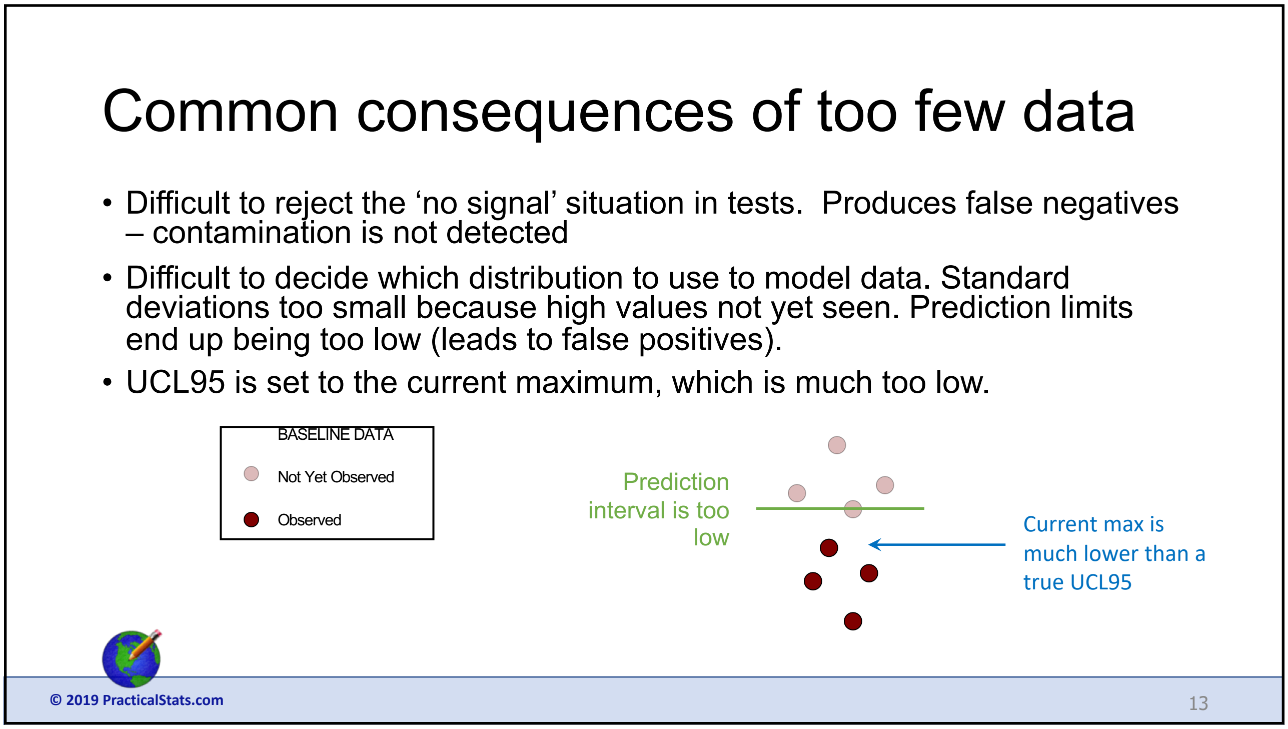
11



ATSDR Guidance to State Health Assessors (soil chemistry; toxics)

- Based on simulation studies
- 8 samples are the minimum to perform any statistical methods. Below that, just plots. With fewer, you don’t know have any idea what the upper and lower percentiles might be. Even estimates of the mean may be far off.
- From 8-20 observations, a distribution will need to be assumed, as this is not enough to be certain that extremes have been sampled
- Above 20 observations, bootstrapping and permutation tests can be performed (not requiring an assumed distribution)

12



Common consequences of too few data

- Difficult to reject the ‘no signal’ situation in tests. Produces false negatives – contamination is not detected
- Difficult to decide which distribution to use to model data. Standard deviations too small because high values not yet seen. Prediction limits end up being too low (leads to false positives).
- UCL95 is set to the current maximum, which is much too low.

BASELINE DATA

○ Not Yet Observed

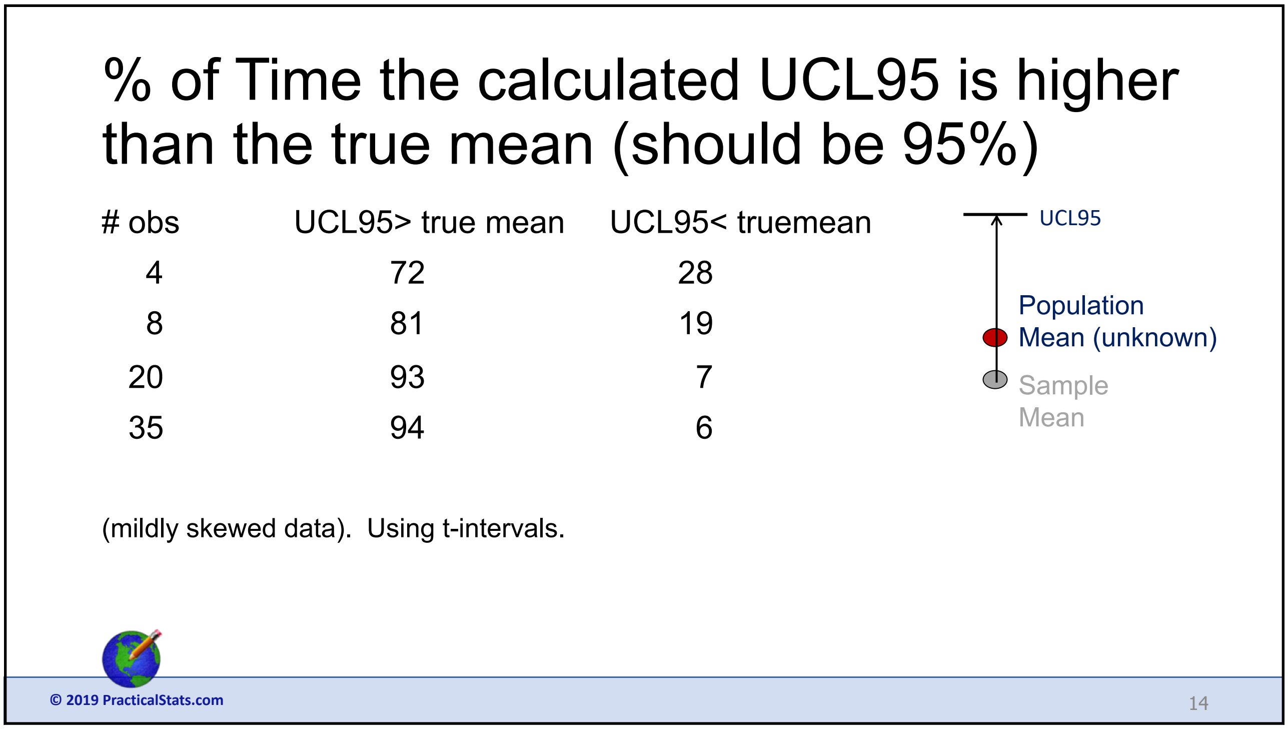
● Observed

Prediction interval is too low

Current max is much lower than a true UCL95

© 2019 PracticalStats.com

13



% of Time the calculated UCL95 is higher than the true mean (should be 95%)

# obs	UCL95> true mean	UCL95< truemean
4	72	28
8	81	19
20	93	7
35	94	6

UCL95

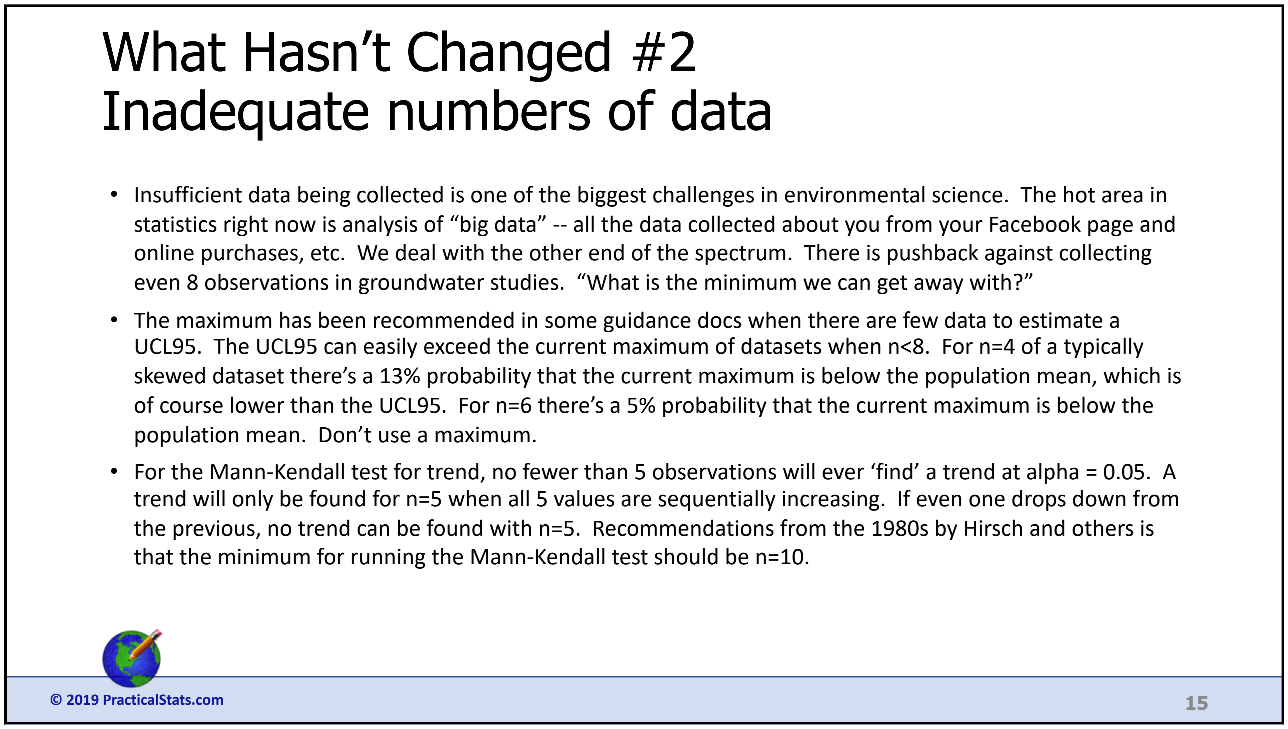
Population Mean (unknown)

Sample Mean

(mildly skewed data). Using t-intervals.

© 2019 PracticalStats.com

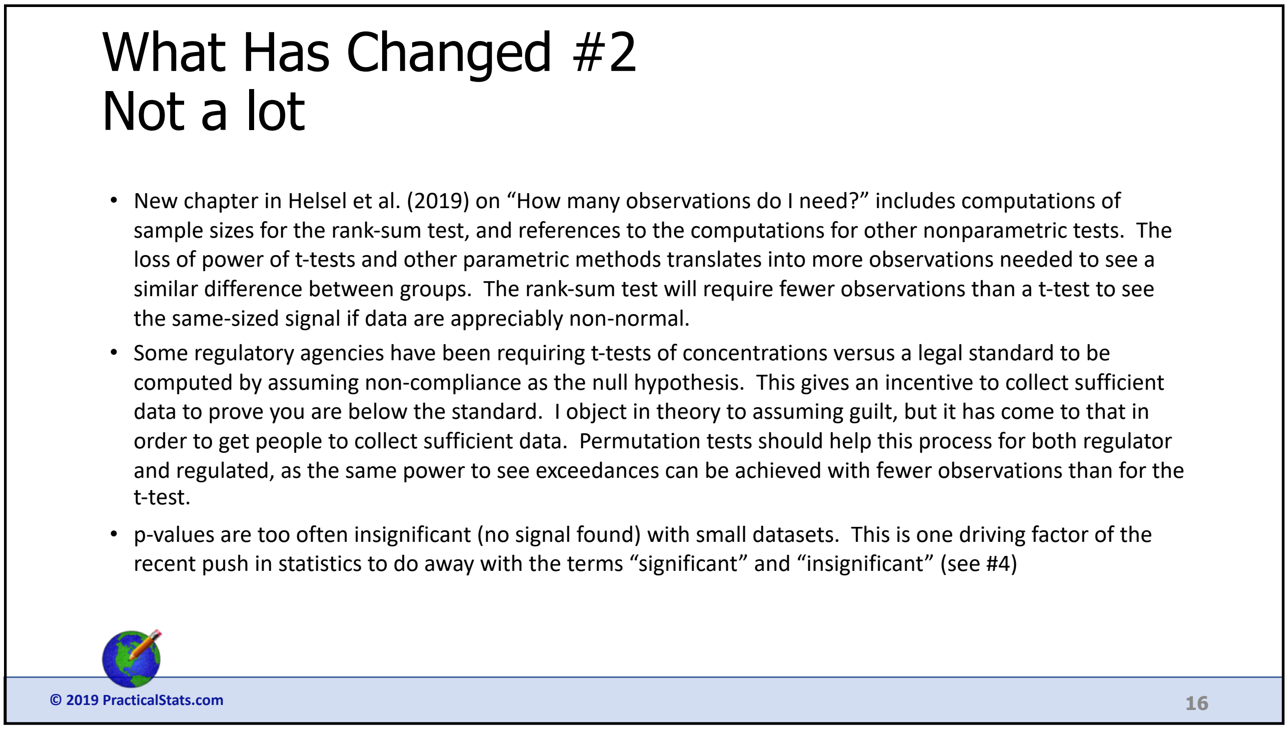
14



What Hasn't Changed #2
Inadequate numbers of data

- Insufficient data being collected is one of the biggest challenges in environmental science. The hot area in statistics right now is analysis of "big data" -- all the data collected about you from your Facebook page and online purchases, etc. We deal with the other end of the spectrum. There is pushback against collecting even 8 observations in groundwater studies. "What is the minimum we can get away with?"
- The maximum has been recommended in some guidance docs when there are few data to estimate a UCL95. The UCL95 can easily exceed the current maximum of datasets when n<8. For n=4 of a typically skewed dataset there's a 13% probability that the current maximum is below the population mean, which is of course lower than the UCL95. For n=6 there's a 5% probability that the current maximum is below the population mean. Don't use a maximum.
- For the Mann-Kendall test for trend, no fewer than 5 observations will ever 'find' a trend at alpha = 0.05. A trend will only be found for n=5 when all 5 values are sequentially increasing. If even one drops down from the previous, no trend can be found with n=5. Recommendations from the 1980s by Hirsch and others is that the minimum for running the Mann-Kendall test should be n=10.

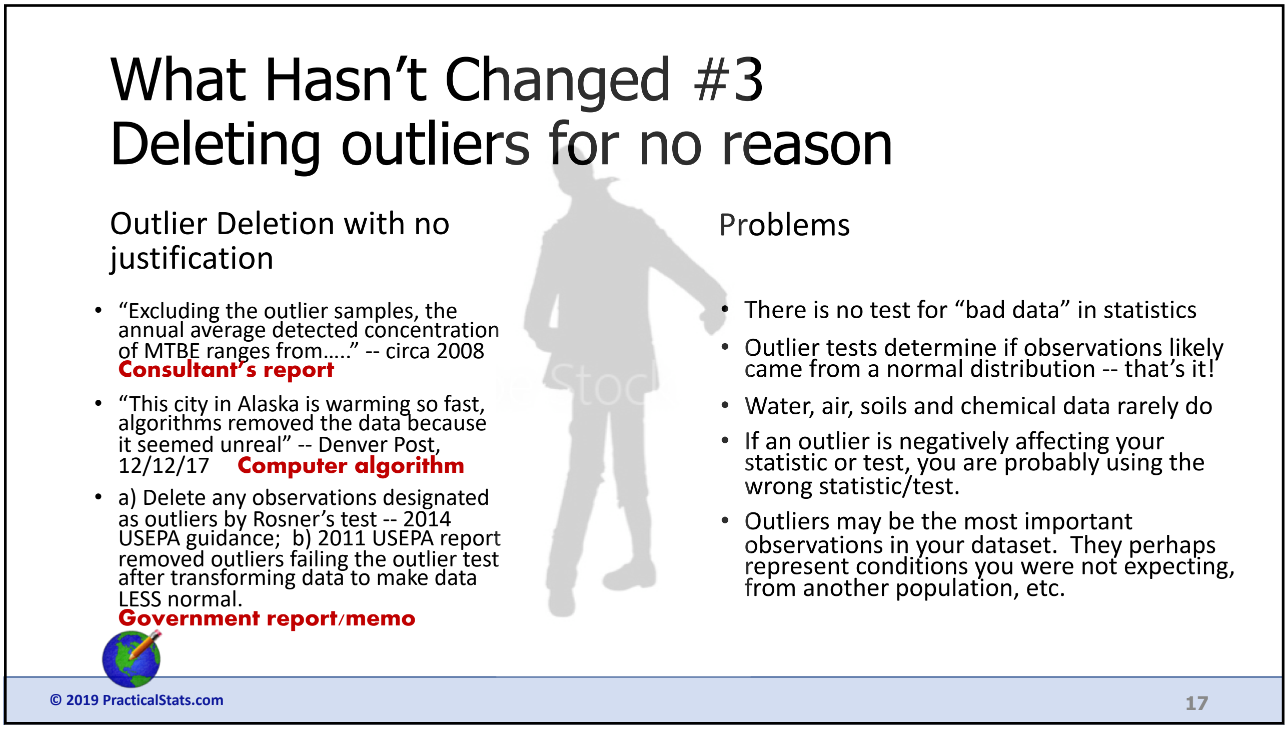
© 2019 PracticalStats.com 15



What Has Changed #2
Not a lot

- New chapter in Helsel et al. (2019) on "How many observations do I need?" includes computations of sample sizes for the rank-sum test, and references to the computations for other nonparametric tests. The loss of power of t-tests and other parametric methods translates into more observations needed to see a similar difference between groups. The rank-sum test will require fewer observations than a t-test to see the same-sized signal if data are appreciably non-normal.
- Some regulatory agencies have been requiring t-tests of concentrations versus a legal standard to be computed by assuming non-compliance as the null hypothesis. This gives an incentive to collect sufficient data to prove you are below the standard. I object in theory to assuming guilt, but it has come to that in order to get people to collect sufficient data. Permutation tests should help this process for both regulator and regulated, as the same power to see exceedances can be achieved with fewer observations than for the t-test.
- p-values are too often insignificant (no signal found) with small datasets. This is one driving factor of the recent push in statistics to do away with the terms "significant" and "insignificant" (see #4)

© 2019 PracticalStats.com 16



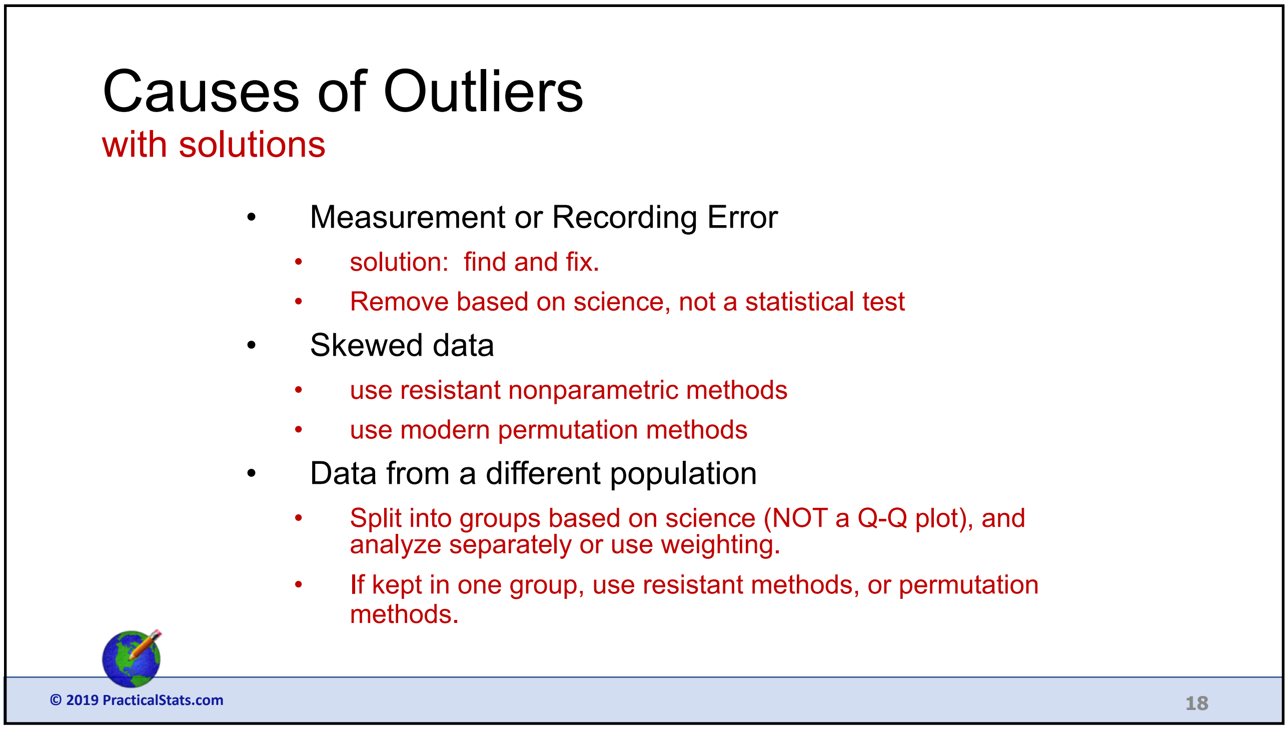
What Hasn’t Changed #3 Deleting outliers for no reason

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 -- that’s it!
- 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.
- Outliers may be the most important observations in your dataset. They perhaps represent conditions you were not expecting, from another population, etc.



Causes of Outliers

with solutions

- Measurement or Recording Error
  - solution: find and fix.
  - Remove based on science, not a statistical test
- Skewed data
  - use resistant nonparametric methods
  - use modern permutation methods
- Data from a different population
  - Split into groups based on science (NOT a Q-Q plot), and analyze separately or use weighting.
  - If kept in one group, use resistant methods, or permutation methods.

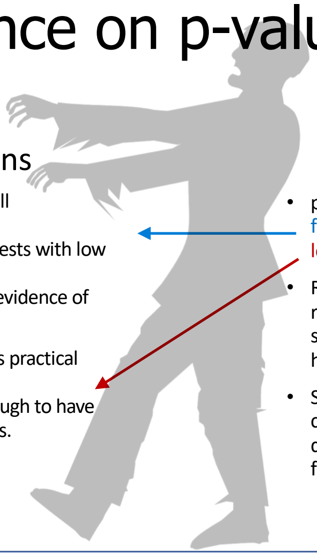
How to think about outliers

- Outliers have a disproportionate effect on the mean.
- If you want a typical value, use the median and don't worry about them.
- If you want an estimate of the mass or total amount or cumulative exposure, they SHOULD effect the result unless they are in error.
- Outliers can be caused by common skewness, and don't necessarily indicate an error

What Has Changed #3
Outlier deletion 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 in court why they deleted them. Was it personal bias to just get what you wanted to see? 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

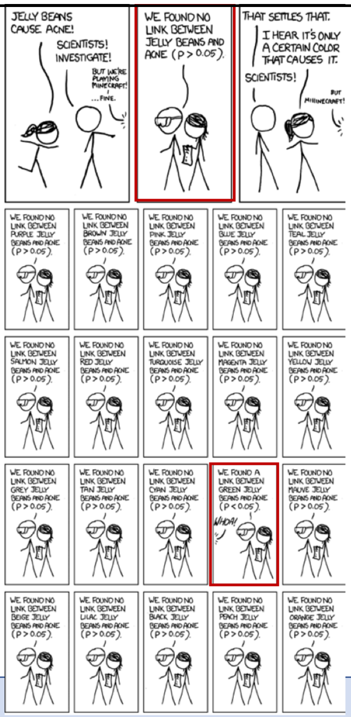


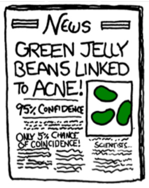
Misconceptions

- ”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

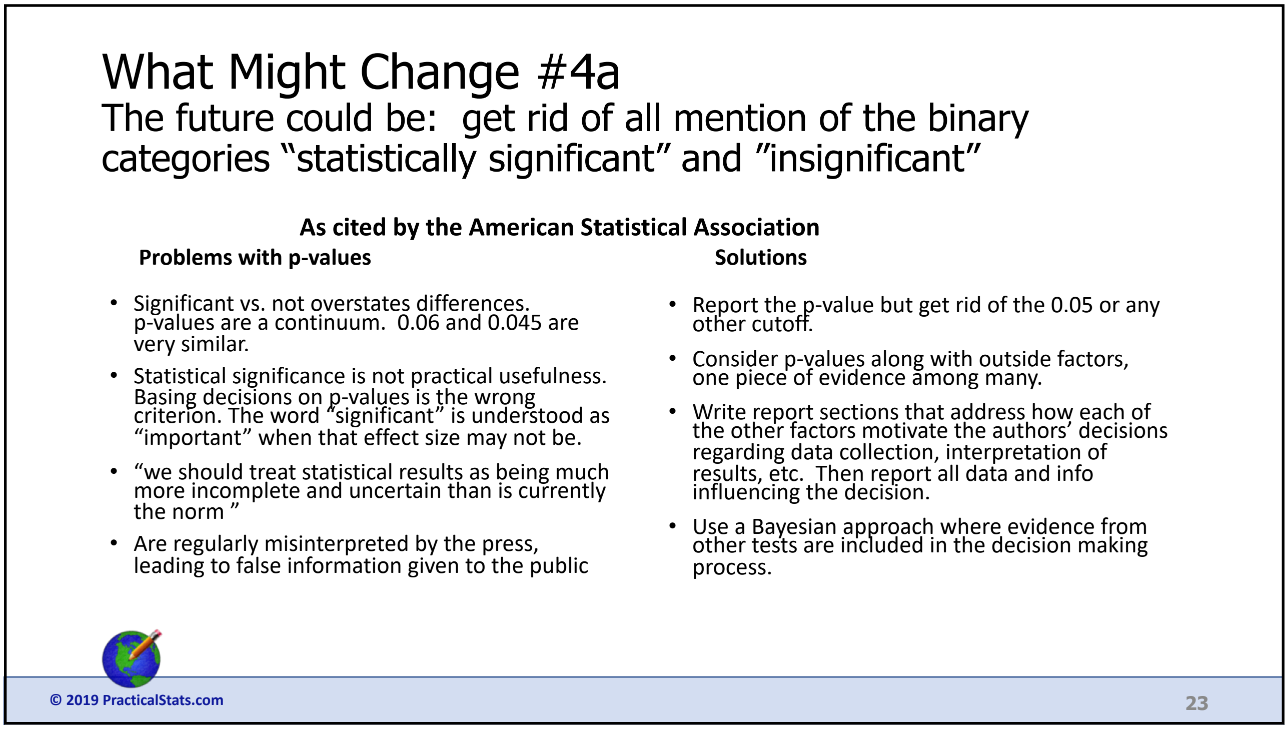
- p-values are a function of sample size.
  few data: large trends may not be seen
  lots of data: unimportant trends found
- Researchers sometimes try multiple hypothesis tests, removal of outliers, deletion of groups, etc. to achieve statistically significant results. This process is termed “p-hacking”.
- 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.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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 Might Change #4a
The future could be: get rid of all mention of the binary categories "statistically significant" and "insignificant"

As cited by the American Statistical Association

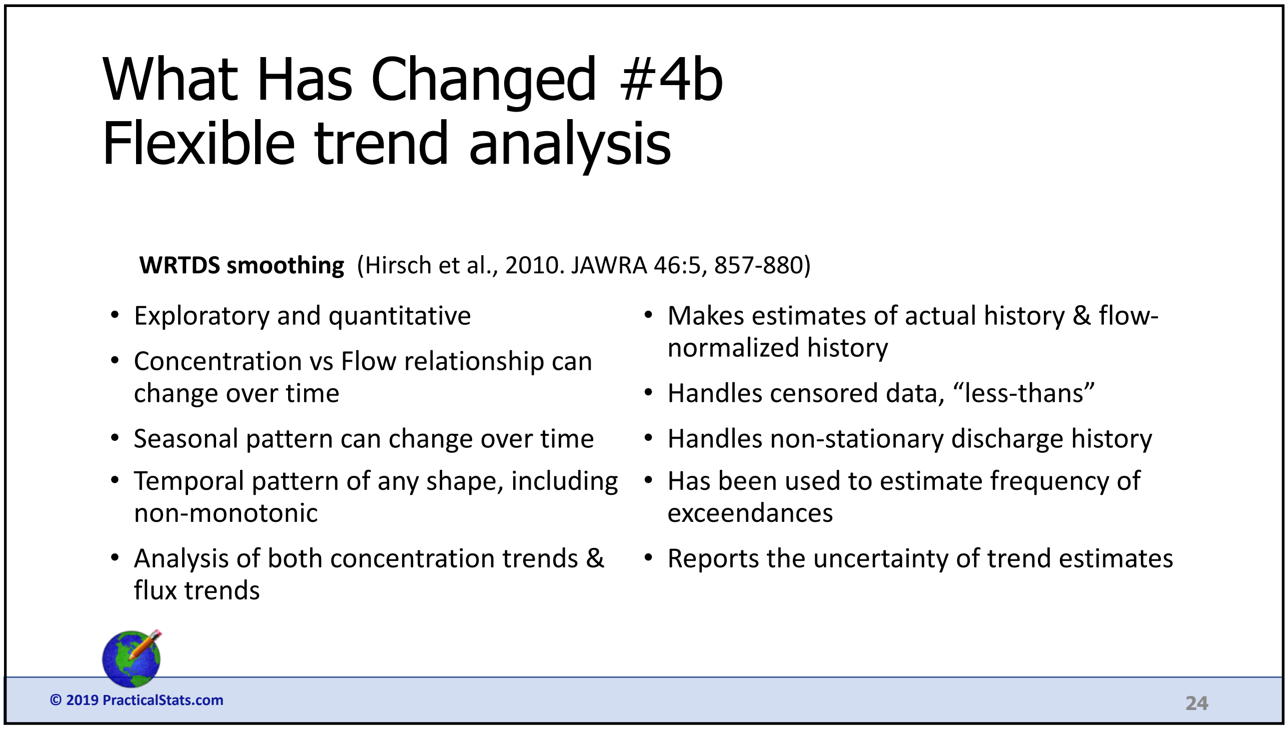
Problems with p-values

- Significant vs. not overstates differences. p-values are a continuum. 0.06 and 0.045 are very similar.
- Statistical significance is not practical usefulness. Basing decisions on p-values is the wrong criterion. The word "significant" is understood as "important" when that effect size may not be.
- "we should treat statistical results as being much more incomplete and uncertain than is currently the norm "
- Are regularly misinterpreted by the press, leading to false information given to the public

Solutions

- Report the p-value but get rid of the 0.05 or any other cutoff.
- Consider p-values along with outside factors, one piece of evidence among many.
- Write report sections that address how each of the other factors motivate the authors' decisions regarding data collection, interpretation of results, etc. Then report all data and info influencing the decision.
- Use a Bayesian approach where evidence from other tests are included in the decision making process.

23



What Has Changed #4b
Flexible trend analysis

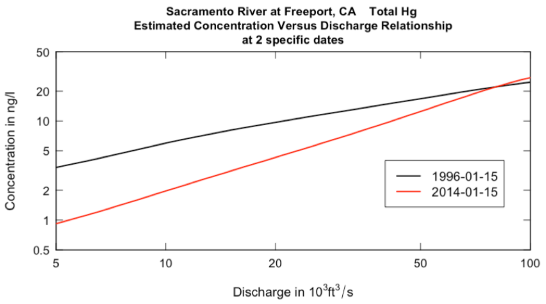
WRTDS smoothing (Hirsch et al., 2010. JAWRA 46:5, 857-880)

- 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

24

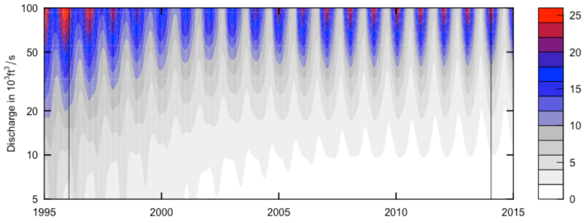
What Has Changed #4b
Flexible trend analysis

Standard regression would force these two lines to be parallel.



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

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

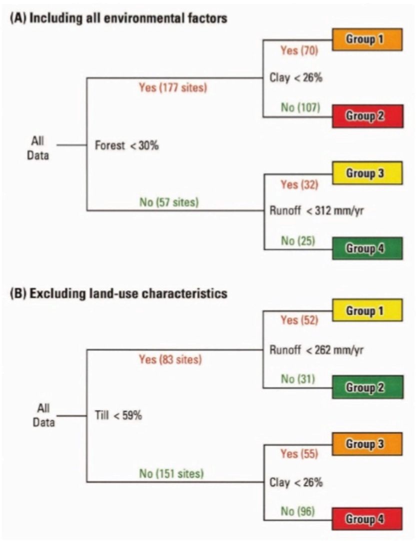


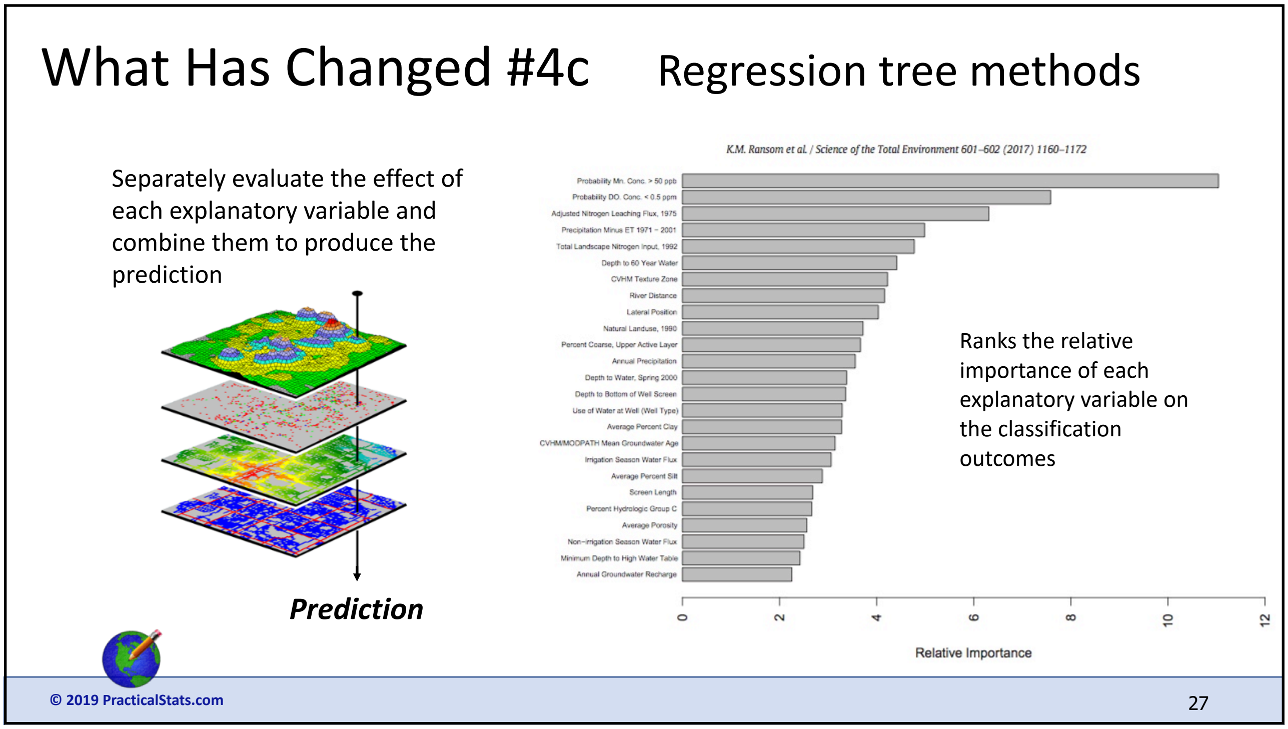
What Has Changed #4c
Regression tree methods

Benefits of Regression Tree methods

- Classifies data into groups by relating the target variable to cutoffs of explanatory variables (Machine Learning)
- Doesn't assume normality or linearity -- very flexible
- Data at the 'high end' do not affect relationships at the 'low end', so not as restricted as are regression methods
- Evaluation of success done by cross-validation, the % of correct predictions of categories for the response variables, rather than by p-values
- Predictions of individual observations rather then for the mean of observations (as done in regression)

594 D. M. Robertson and D. A. Saad





What Has Changed #4c Regression tree methods

Separately evaluate the effect of each explanatory variable and combine them to produce the prediction

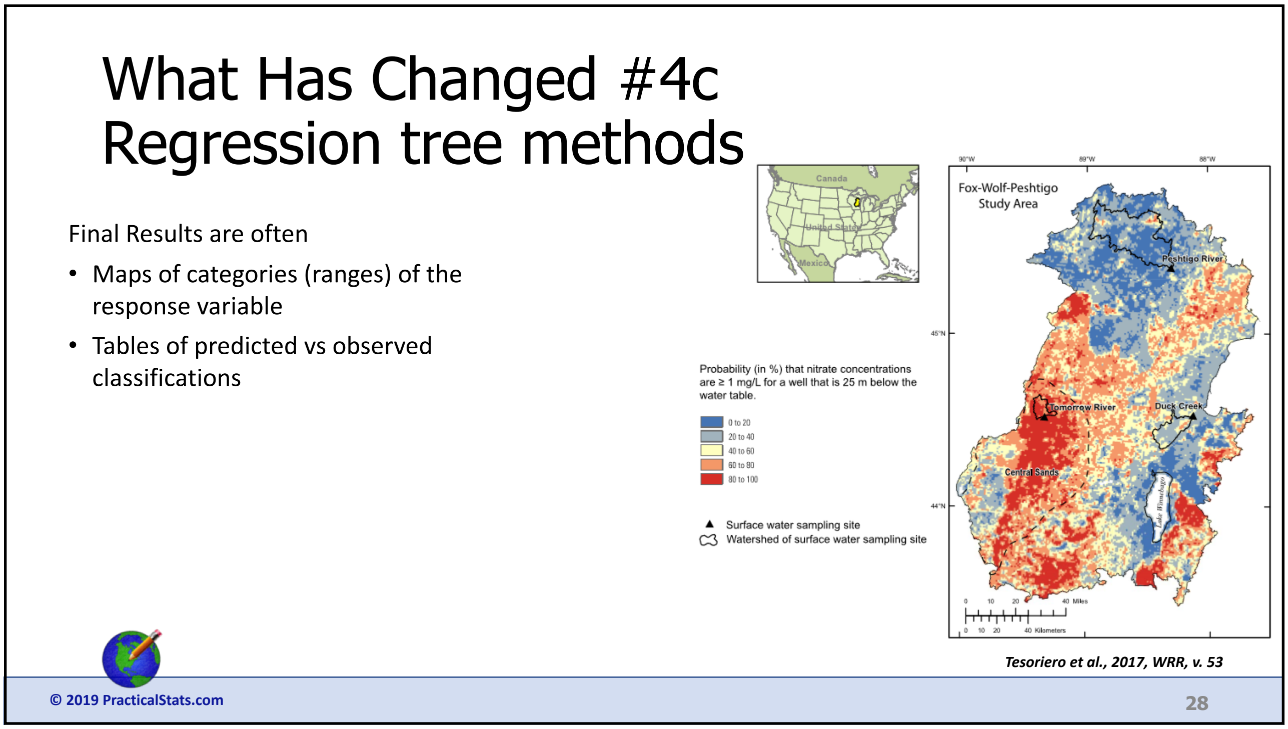
Prediction

Ranks the relative importance of each explanatory variable on the classification outcomes

K.M. Ransom et al. / Science of the Total Environment 601–602 (2017) 1160–1172

© 2019 PracticalStats.com

27



What Has Changed #4c Regression tree methods

Final Results are often

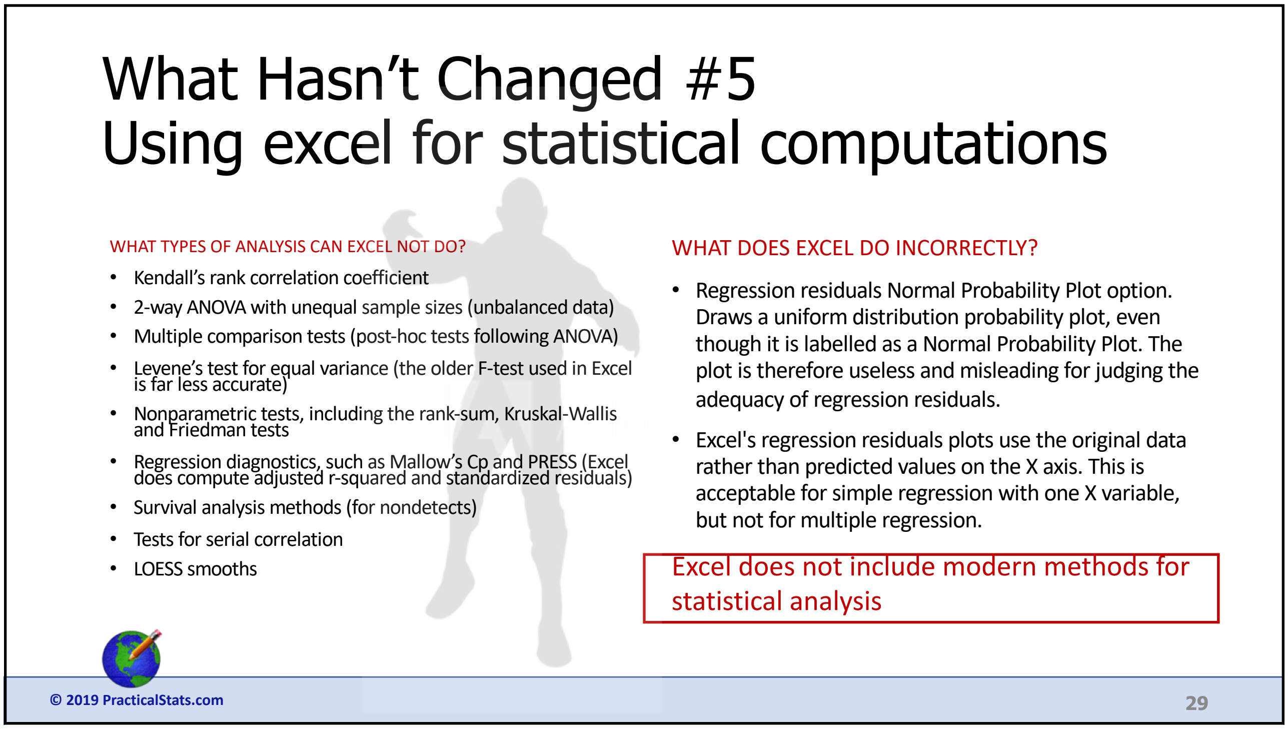
- Maps of categories (ranges) of the response variable
- Tables of predicted vs observed classifications

Probability (in %) that nitrate concentrations are ≥ 1 mg/L for a well that is 25 m below the water table.

Tesoriero et al., 2017, WRR, v. 53

© 2019 PracticalStats.com

28



What Hasn't Changed #5
Using excel for statistical computations

WHAT TYPES OF ANALYSIS CAN EXCEL NOT DO?

- Kendall's rank correlation coefficient
- 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

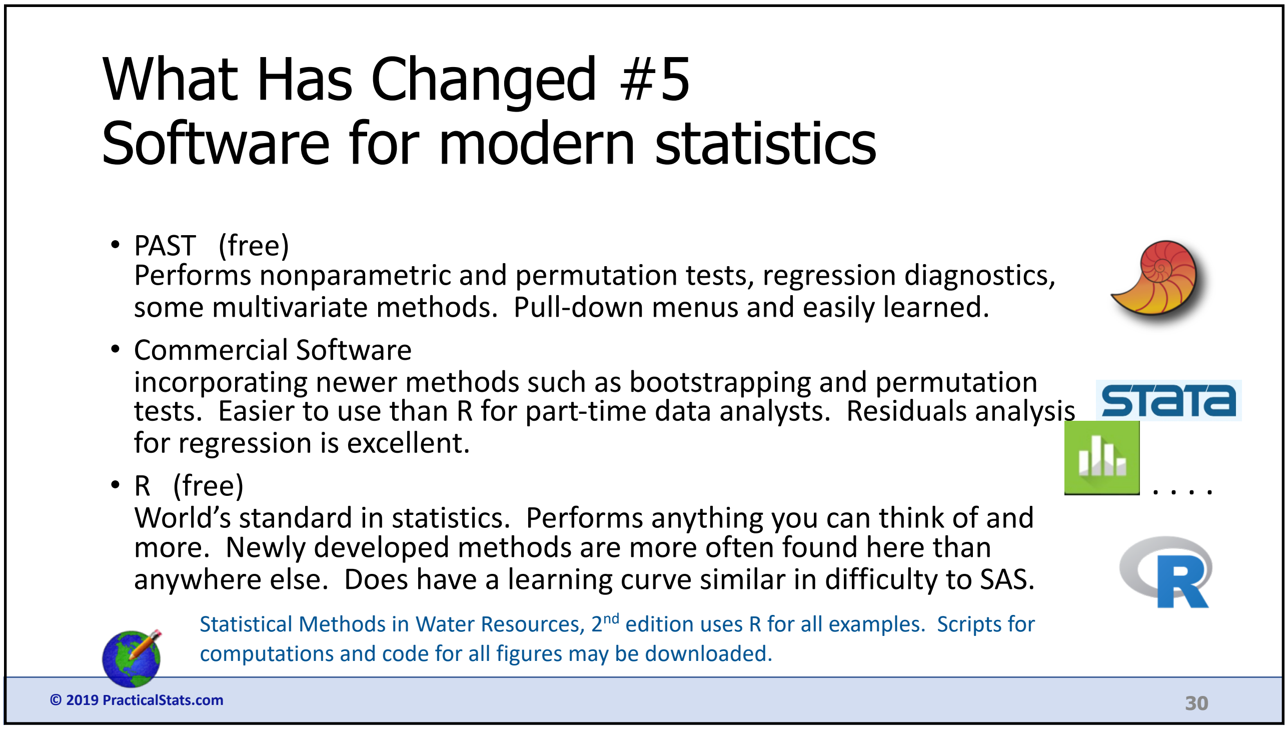
WHAT DOES EXCEL DO INCORRECTLY?

- 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

© 2019 PracticalStats.com

29



What Has 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.

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

© 2019 PracticalStats.com

30

Summary

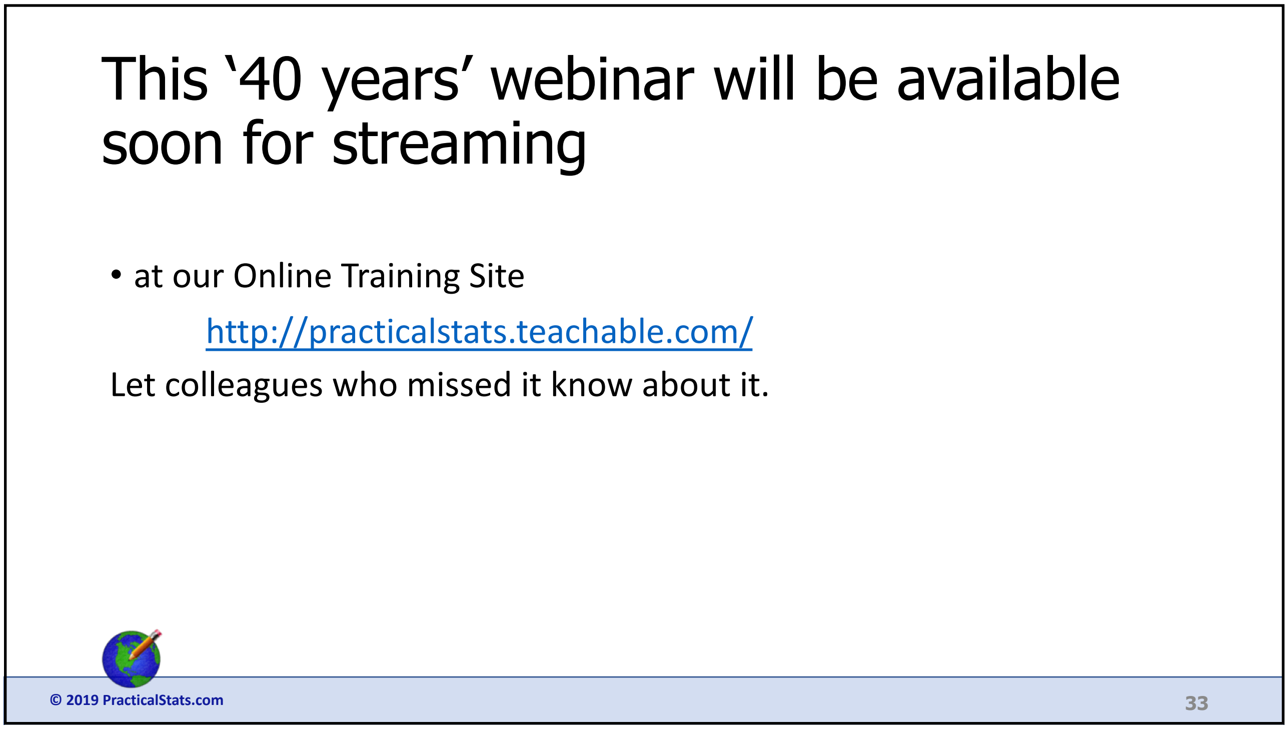
1. Use permutation tests and bootstrapping. You will miss signals if you continue to use old parametric tests and confidence intervals.
2. Collect ample data.
3. Do not delete outliers unless you have evidence outside of the dataset showing they are in error or from another population.
4. Don't "p-hack" or overly rely on p-values. Understand your data using graphical procedures and economic or other data to comprehend and explain the full story.
5. Use modern statistical software. If R seems too complex, try PAST.

Download the second edition of Statistical Methods in Water Resources when it becomes available in 2019 !! Sign up for our webinar/newsletter announcement list to receive a notice when the book is available.

Next Month's Webinar

Tuesday May 21st 11 am Mountain time

- Topic TBD.
- Sign up for our newsletter/announcement list to get the registration link emailed to you. Respond to the survey you'll get in a few minutes to opt into the list, or send email to ask@practicalstats.com
- Or check our webinars page periodically at http://practicalstats.com/training/webinar.html to register for it.

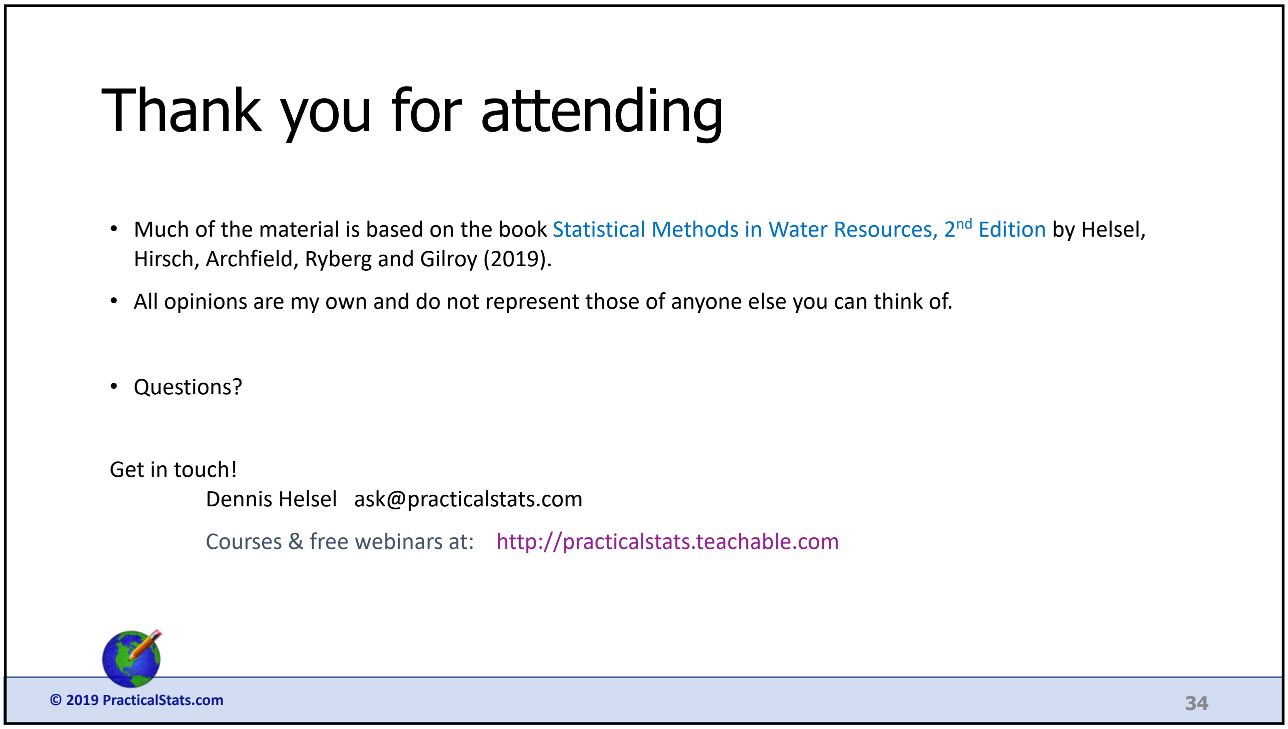


This ‘40 years’ webinar will be available soon for streaming

- at our Online Training Site

  http://practicalstats.teachable.com/

Let colleagues who missed it know about it.



Thank you for attending

- Much of the material is based on the book Statistical Methods in Water Resources, 2nd Edition 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 ask@practicalstats.com
Courses & free webinars at: http://practicalstats.teachable.com