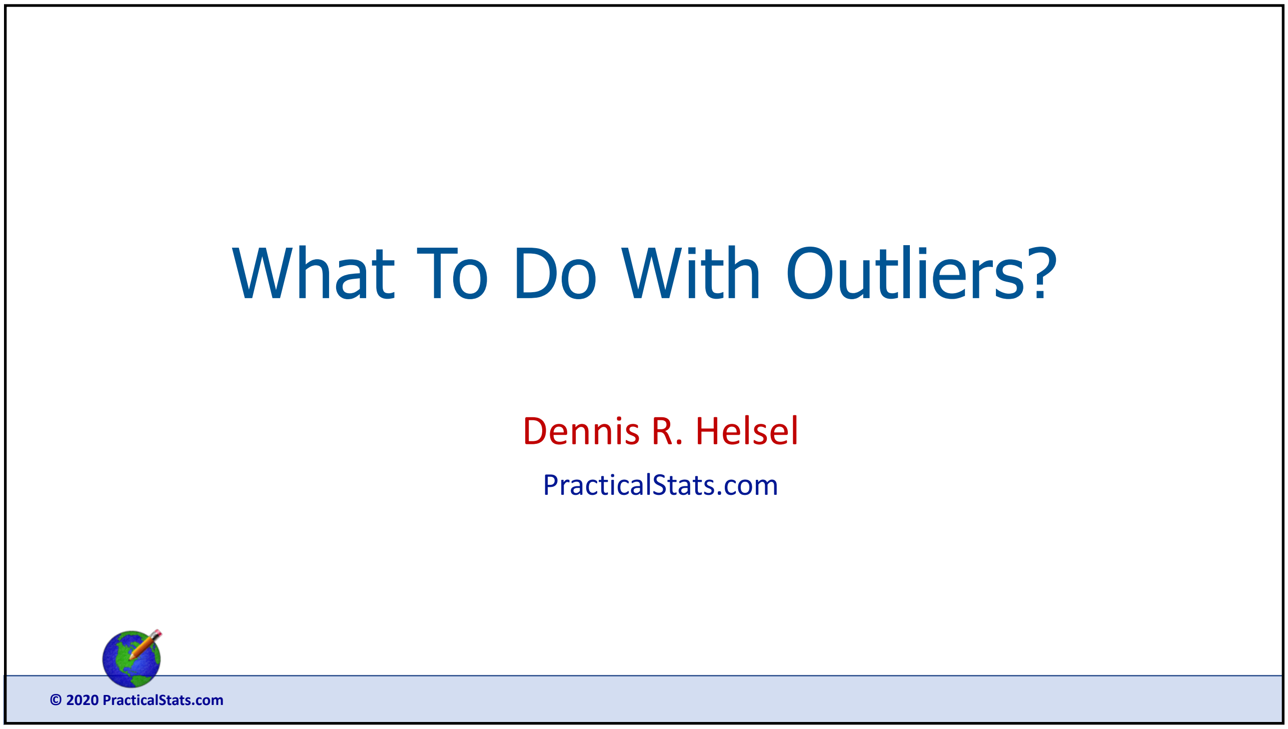


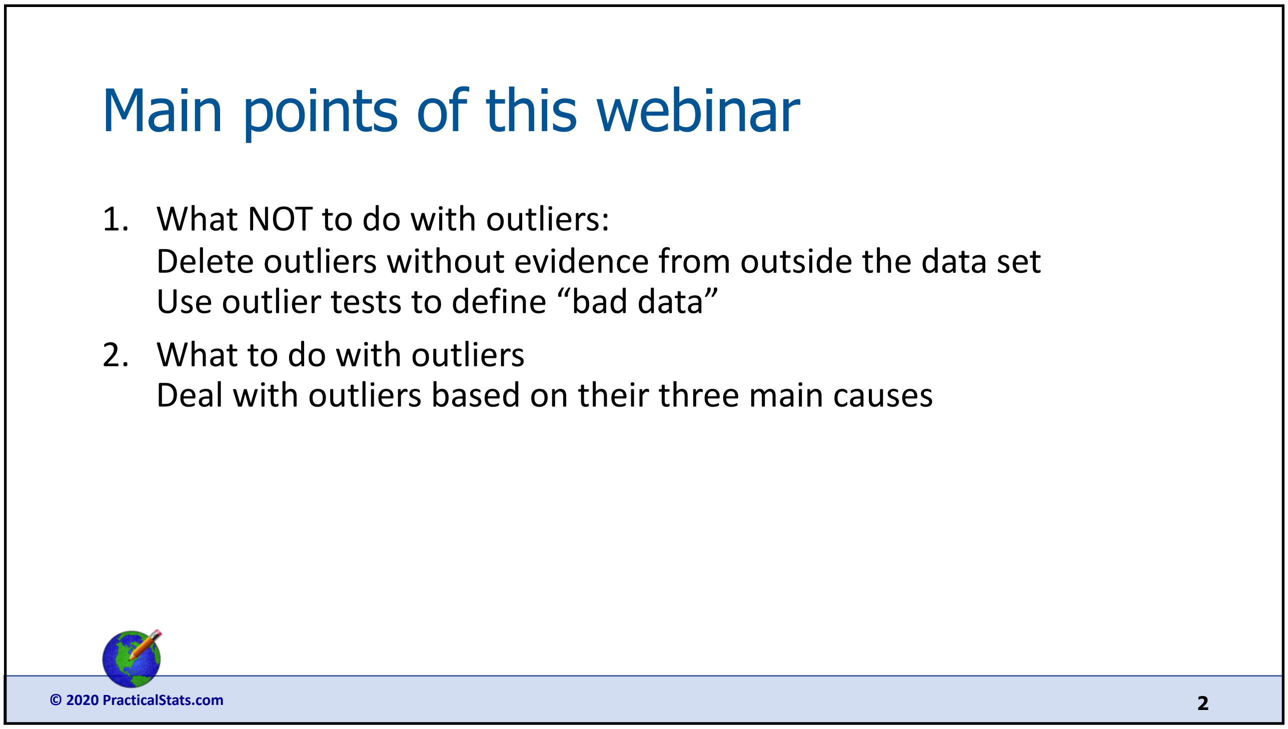
What To Do With Outliers?

Dennis R. Helsel

PracticalStats.com

© 2020 PracticalStats.com

1



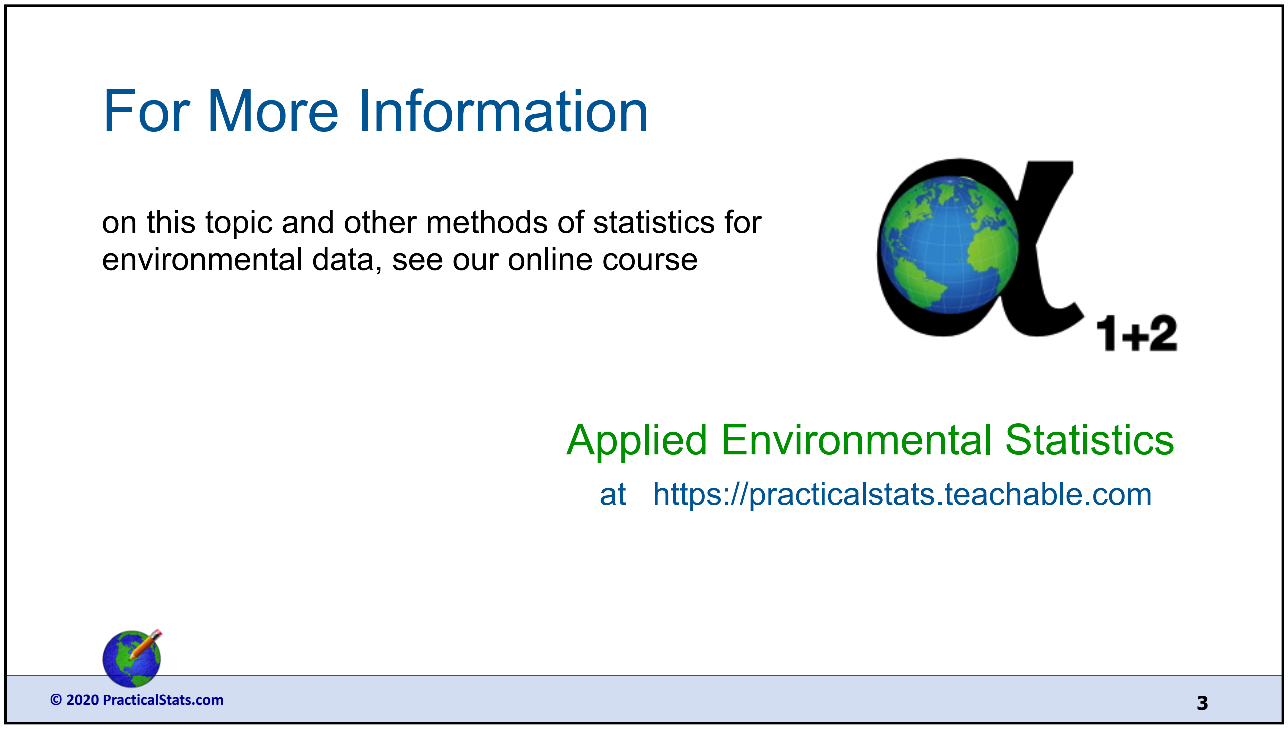
Main points of this webinar

1. What NOT to do with outliers:
   Delete outliers without evidence from outside the data set
   Use outlier tests to define "bad data"
2. What to do with outliers
   Deal with outliers based on their three main causes

© 2020 PracticalStats.com

2

2



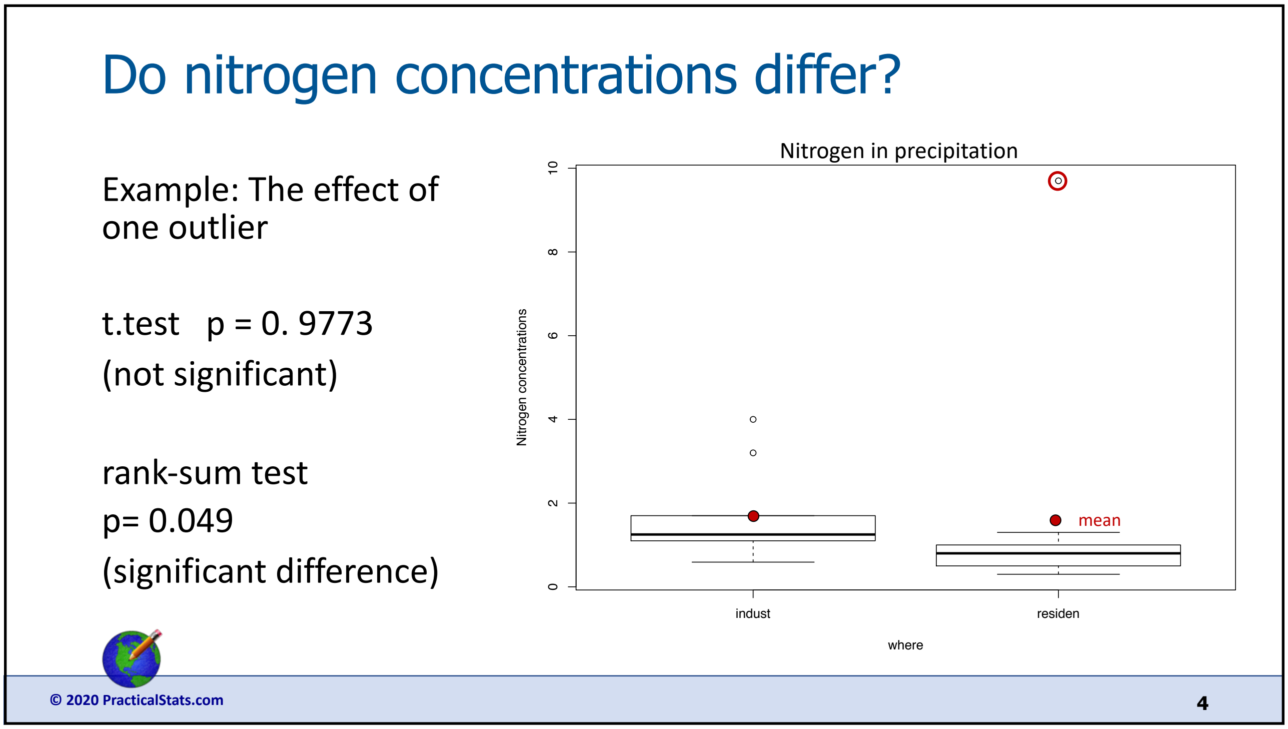
For More Information

on this topic and other methods of statistics for environmental data, see our online course

Applied Environmental Statistics

at https://practicalstats.teachable.com

© 2020 PracticalStats.com



Do nitrogen concentrations differ?

Example: The effect of one outlier

t.test p = 0. 9773

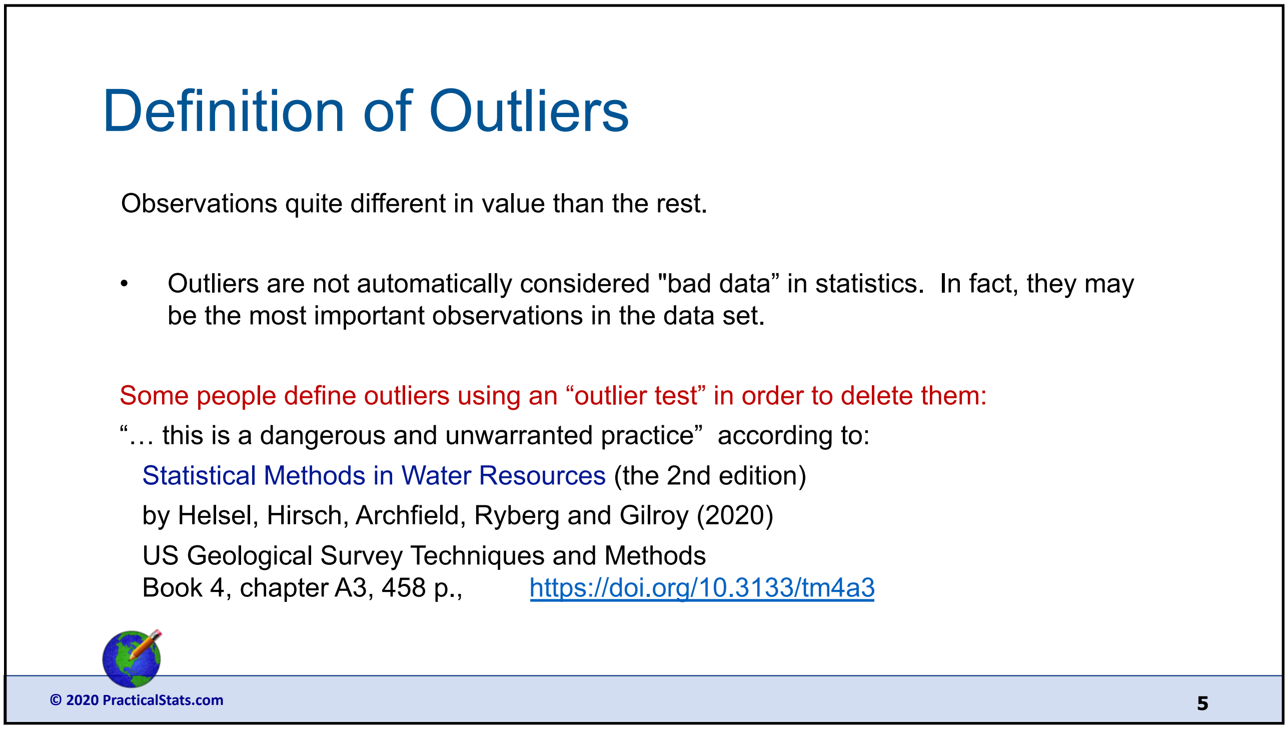
(not significant)

rank-sum test

p= 0.049

(significant difference)

© 2020 PracticalStats.com



Definition of Outliers

Observations quite different in value than the rest.

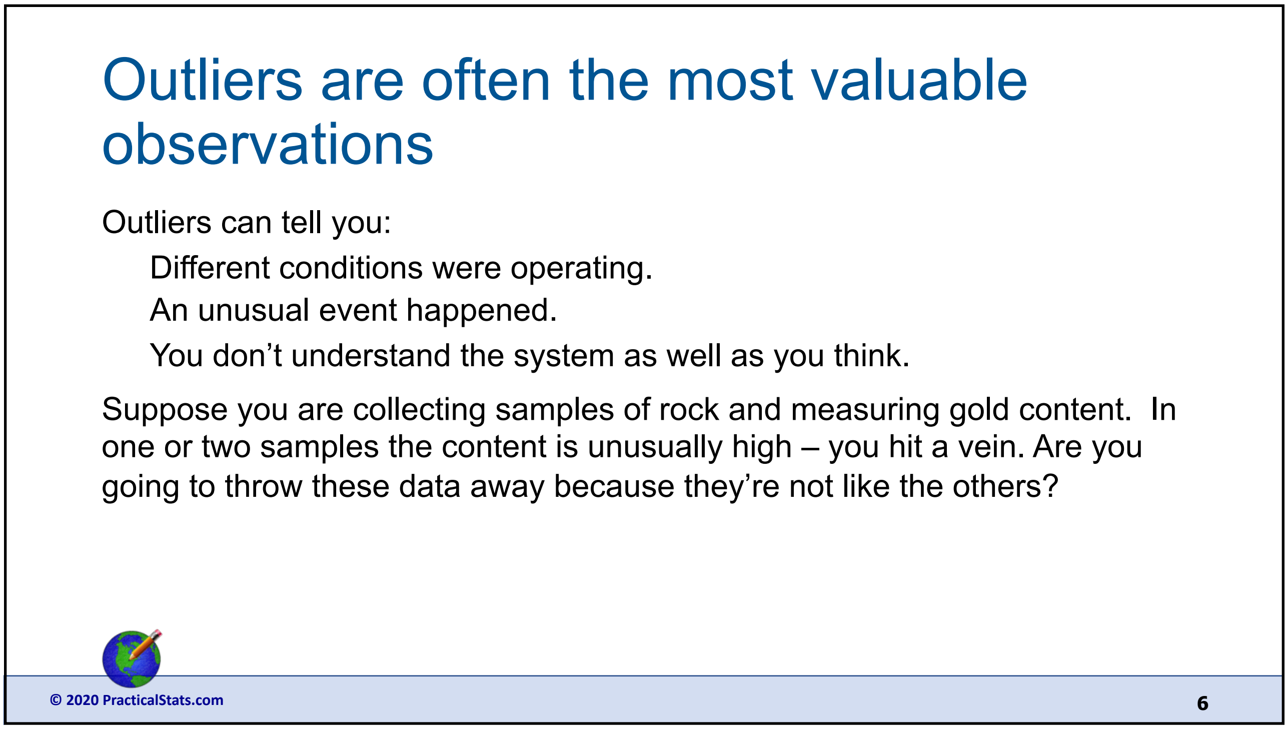
- Outliers are not automatically considered "bad data" in statistics. In fact, they may be the most important observations in the data set.

Some people define outliers using an "outlier test" in order to delete them:

"… this is a dangerous and unwarranted practice" according to:

Statistical Methods in Water Resources (the 2nd edition)
by Helsel, Hirsch, Archfield, Ryberg and Gilroy (2020)
US Geological Survey Techniques and Methods
Book 4, chapter A3, 458 p., https://doi.org/10.3133/tm4a3

© 2020 PracticalStats.com



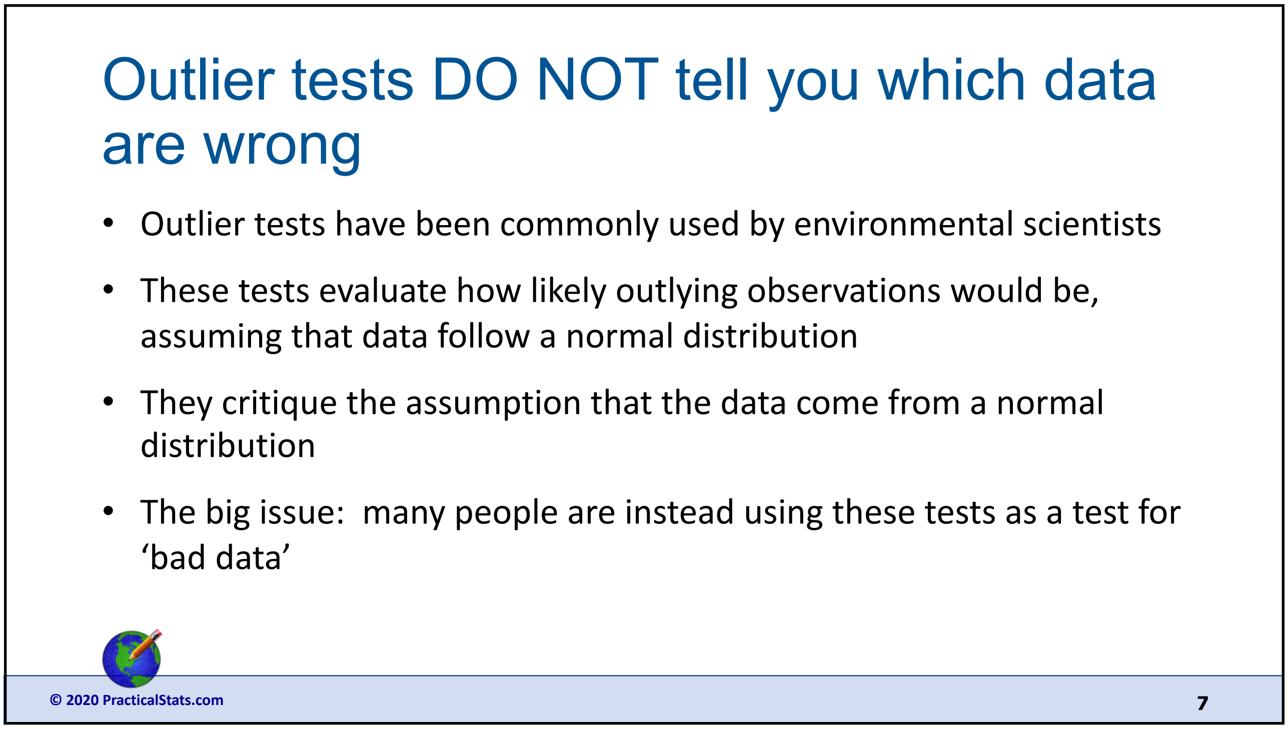
Outliers are often the most valuable observations

Outliers can tell you:

Different conditions were operating.
An unusual event happened.
You don't understand the system as well as you think.

Suppose you are collecting samples of rock and measuring gold content. In one or two samples the content is unusually high – you hit a vein. Are you going to throw these data away because they're not like the others?

© 2020 PracticalStats.com



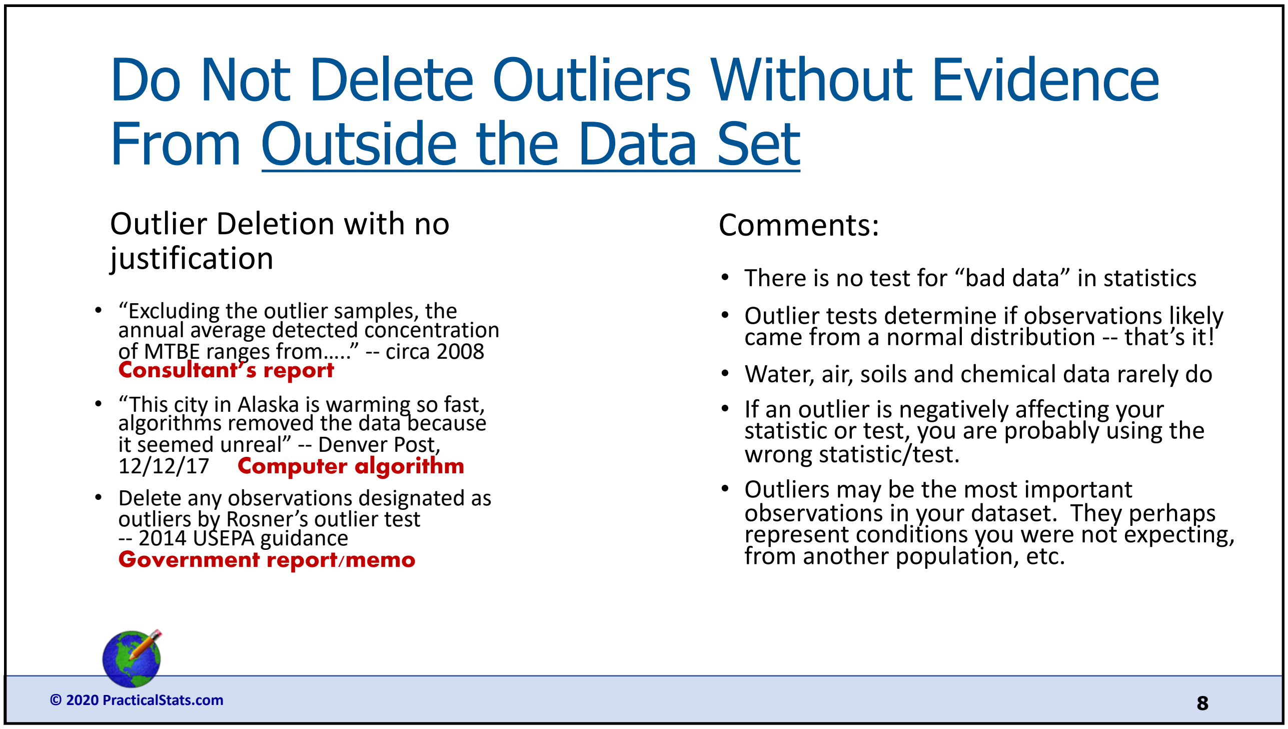
Outlier tests DO NOT tell you which data are wrong

- Outlier tests have been commonly used by environmental scientists
- These tests evaluate how likely outlying observations would be, assuming that data follow a normal distribution
- They critique the assumption that the data come from a normal distribution
- The big issue: many people are instead using these tests as a test for ‘bad data’

© 2020 PracticalStats.com

7

7



Do Not Delete Outliers Without Evidence From Outside the Data Set

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**
- Delete any observations designated as outliers by Rosner’s outlier test -- 2014 USEPA guidance **Government report/memo**

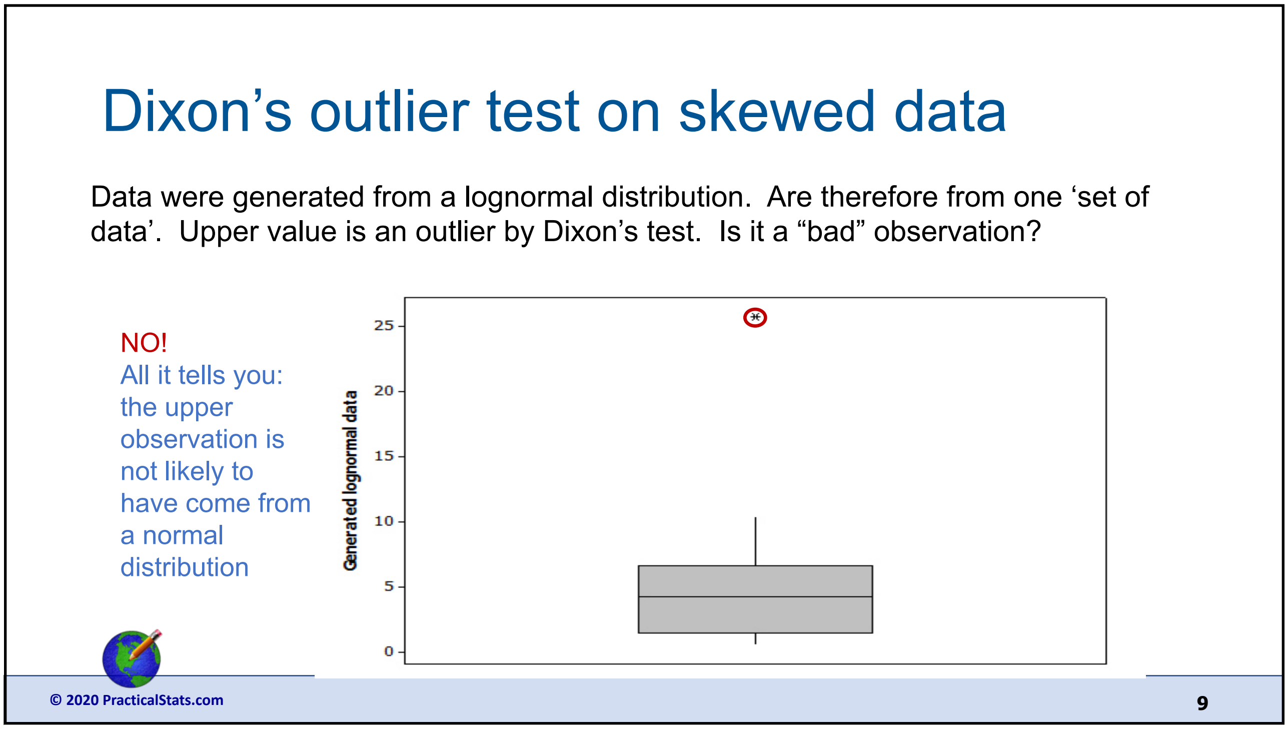
Comments:

- 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.

© 2020 PracticalStats.com

8

8

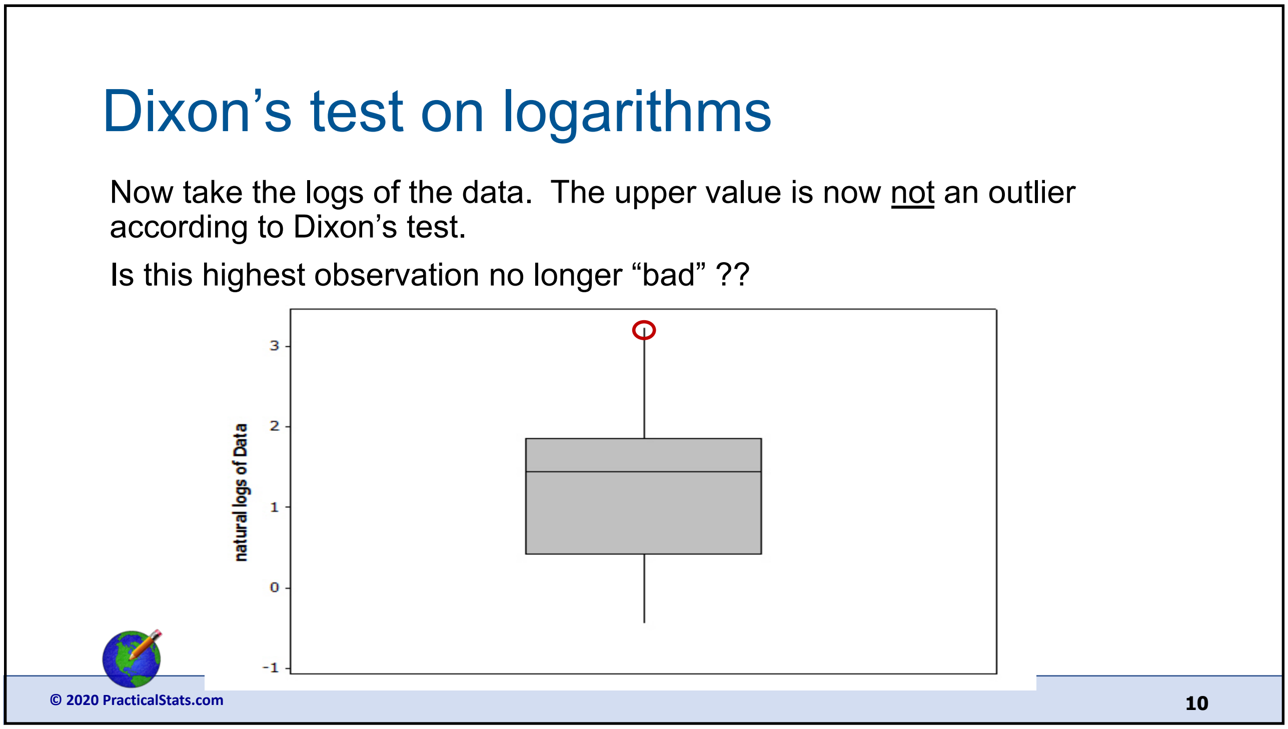


Dixon’s outlier test on skewed data

Data were generated from a lognormal distribution. Are therefore from one ‘set of data’. Upper value is an outlier by Dixon’s test. Is it a “bad” observation?

NO!
All it tells you: the upper observation is not likely to have come from a normal distribution

© 2020 PracticalStats.com

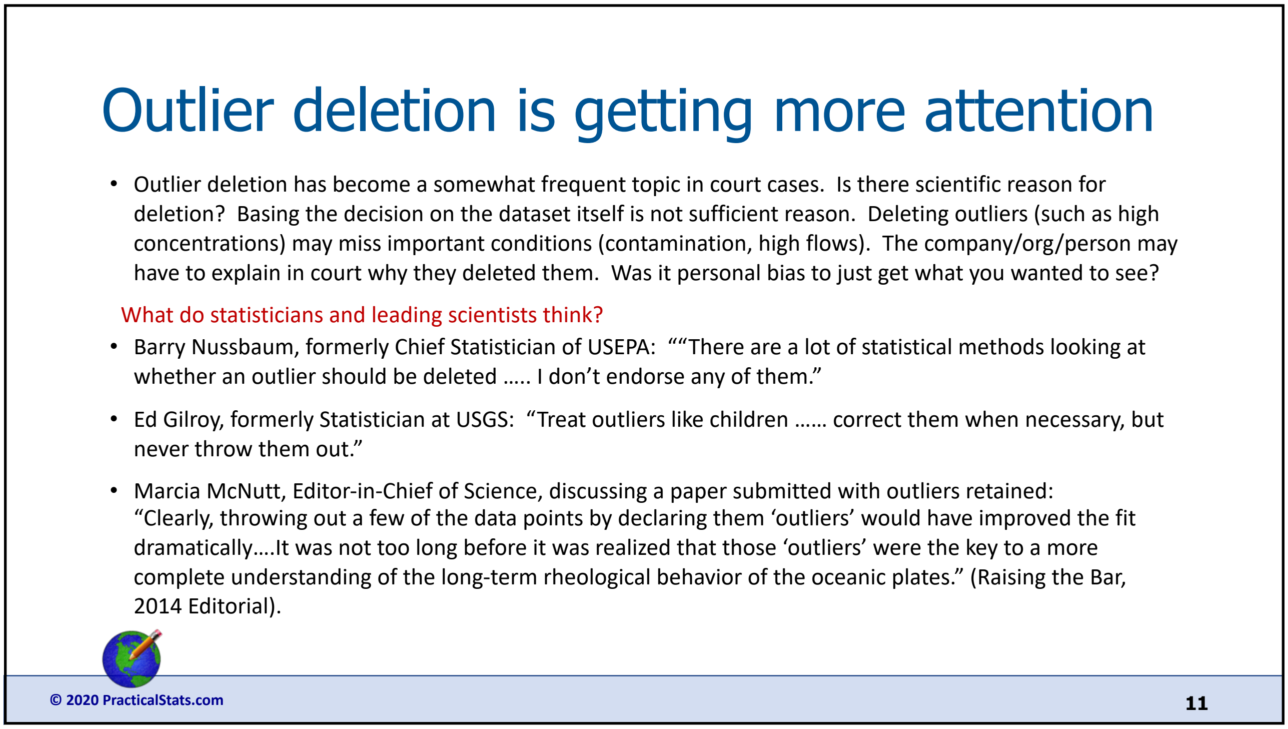


Dixon’s test on logarithms

Now take the logs of the data. The upper value is now not an outlier according to Dixon’s test.

Is this highest observation no longer “bad” ??

© 2020 PracticalStats.com



Outlier deletion is getting more attention

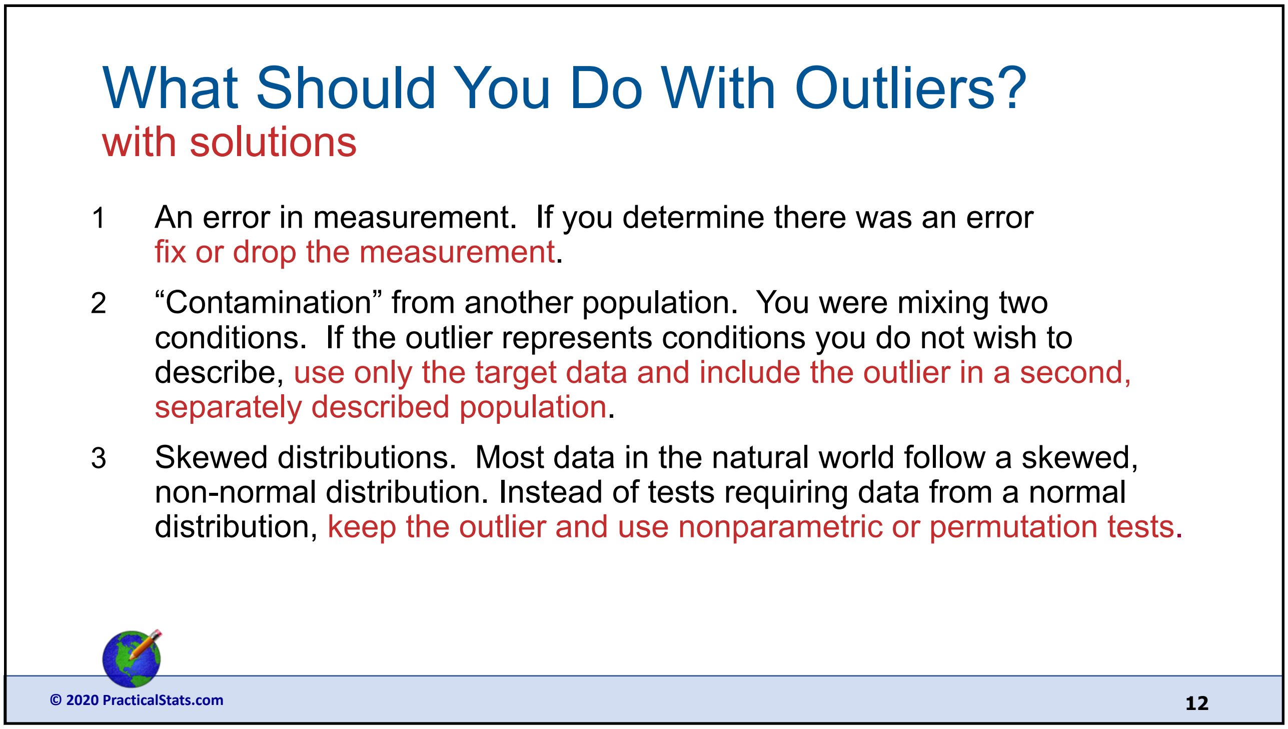
- Outlier deletion has become a somewhat frequent topic in court cases. Is there scientific reason for deletion? Basing the decision on the dataset itself is not sufficient reason. Deleting outliers (such as high concentrations) may miss important conditions (contamination, high flows). The company/org/person 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, discussing a paper submitted with outliers retained: “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).

© 2020 PracticalStats.com 11

11



What Should You Do With Outliers?
with solutions

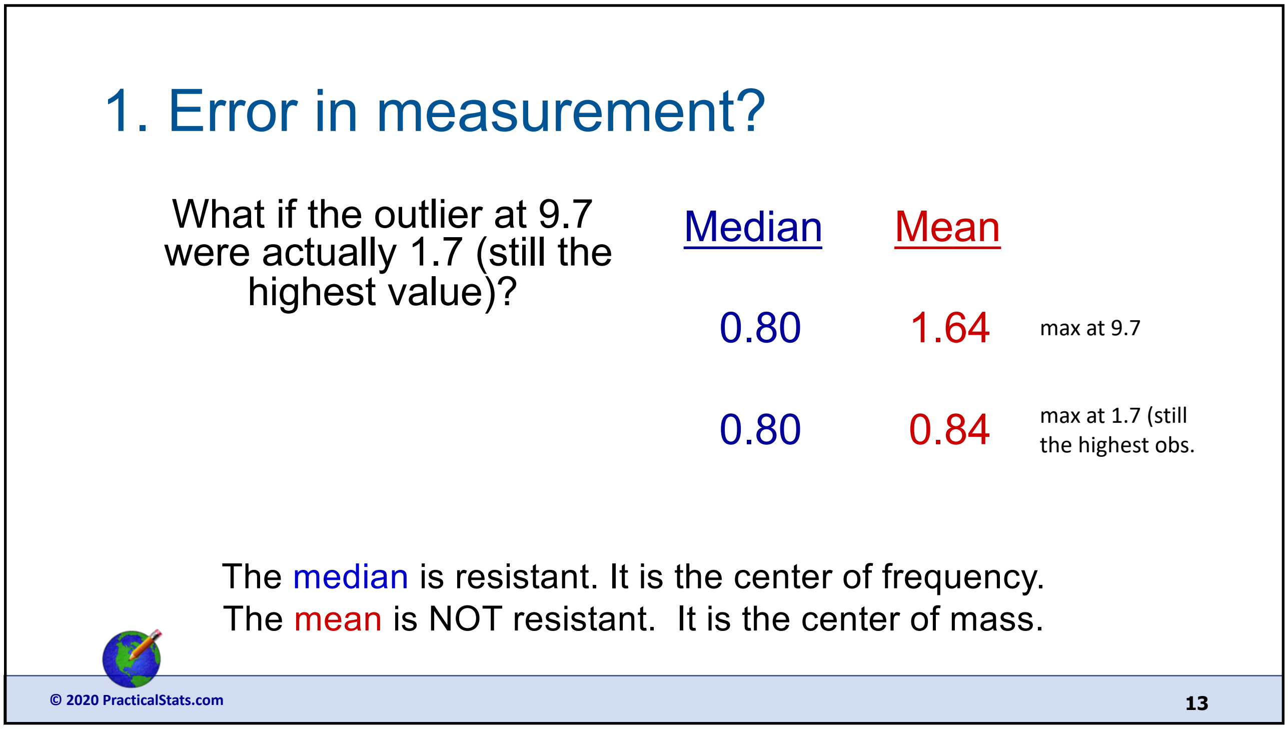
1 An error in measurement. If you determine there was an error fix or drop the measurement.

2 “Contamination” from another population. You were mixing two conditions. If the outlier represents conditions you do not wish to describe, use only the target data and include the outlier in a second, separately described population.

3 Skewed distributions. Most data in the natural world follow a skewed, non-normal distribution. Instead of tests requiring data from a normal distribution, keep the outlier and use nonparametric or permutation tests.

© 2020 PracticalStats.com 12

12



1. Error in measurement?

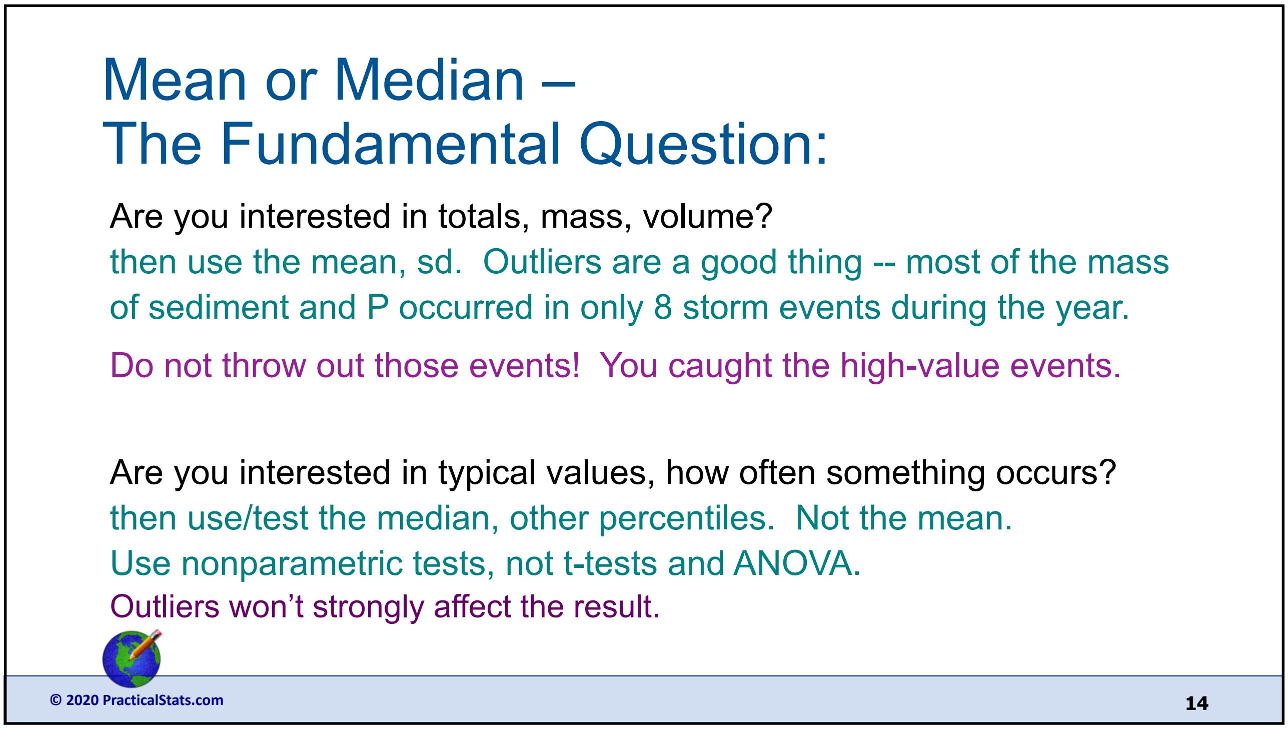
What if the outlier at 9.7 were actually 1.7 (still the highest value)?

Median	Mean	
0.80	1.64	max at 9.7
0.80	0.84	max at 1.7 (still the highest obs.

The median is resistant. It is the center of frequency.
The mean is NOT resistant. It is the center of mass.

© 2020 PracticalStats.com 13

13



Mean or Median – The Fundamental Question:

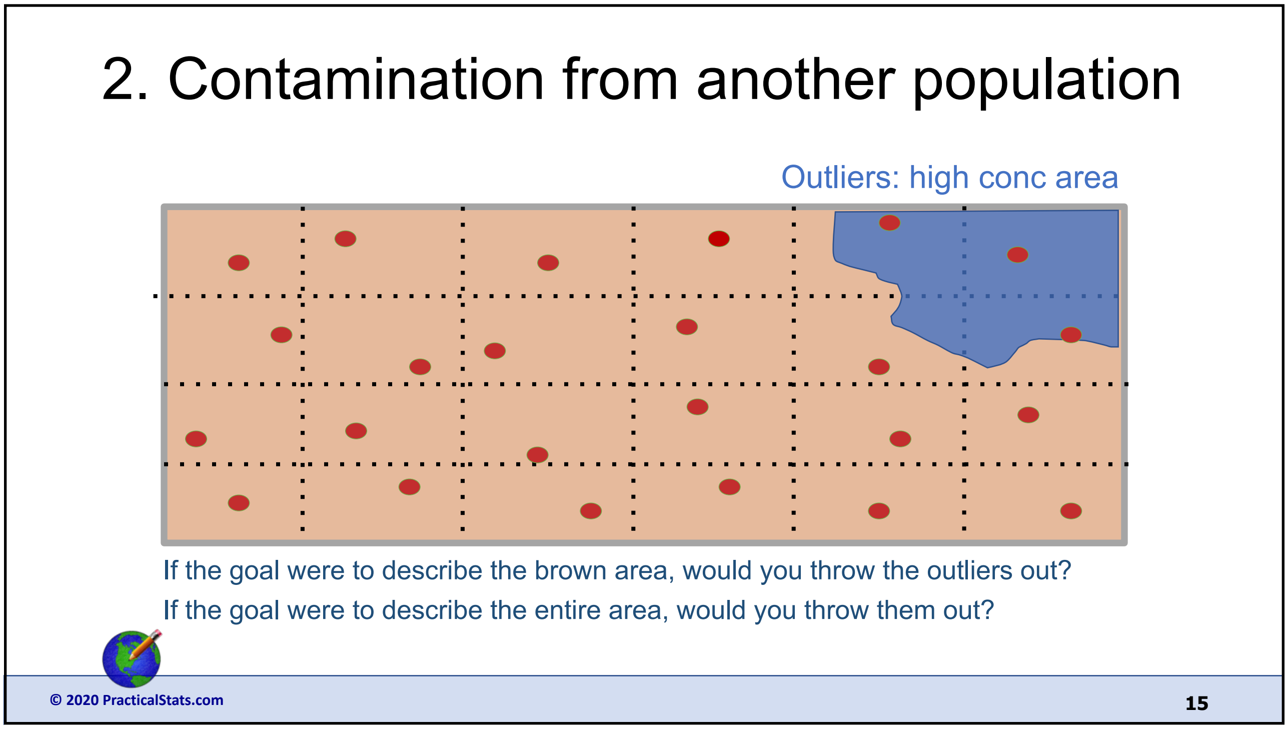
Are you interested in totals, mass, volume?
then use the mean, sd. Outliers are a good thing -- most of the mass of sediment and P occurred in only 8 storm events during the year.

Do not throw out those events! You caught the high-value events.

Are you interested in typical values, how often something occurs?
then use/test the median, other percentiles. Not the mean.
Use nonparametric tests, not t-tests and ANOVA.
Outliers won’t strongly affect the result.

© 2020 PracticalStats.com 14

14



2. Contamination from another population

Outliers: high conc area

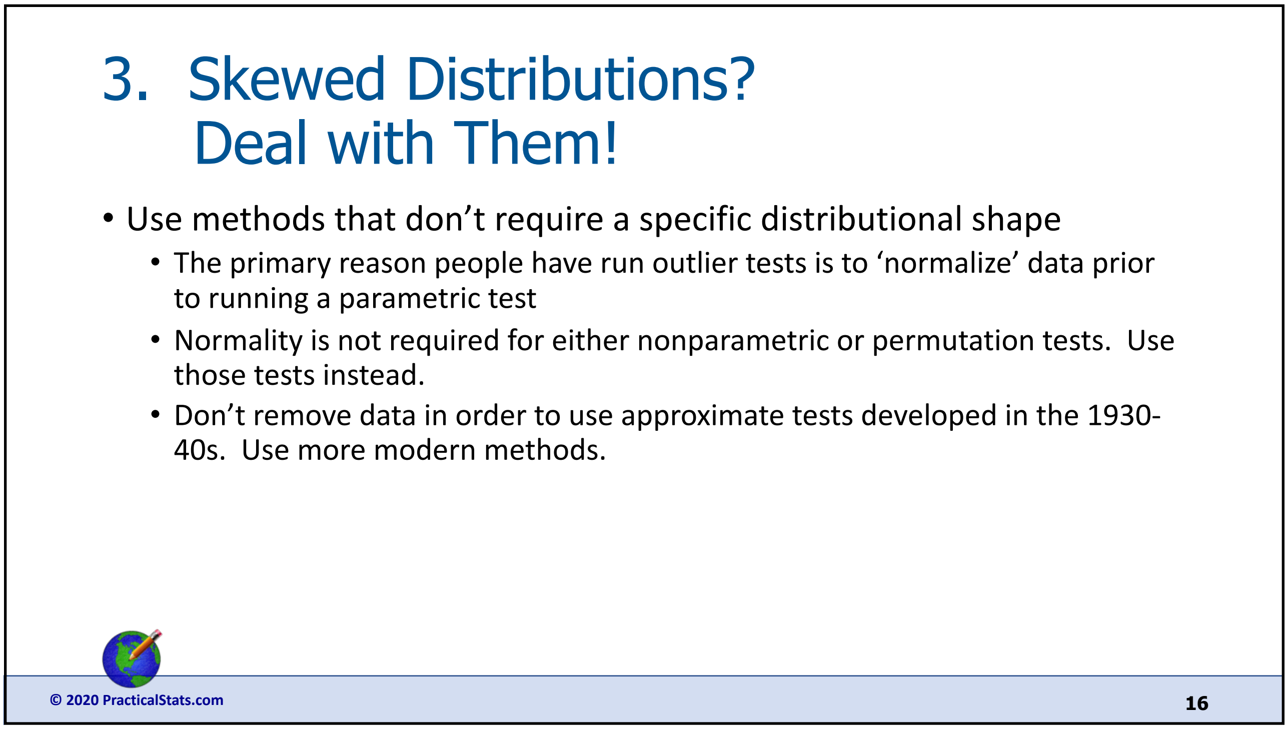
If the goal were to describe the brown area, would you throw the outliers out?

If the goal were to describe the entire area, would you throw them out?

© 2020 PracticalStats.com

15

15



3. Skewed Distributions? Deal with Them!

- Use methods that don't require a specific distributional shape
  - The primary reason people have run outlier tests is to 'normalize' data prior to running a parametric test
  - Normality is not required for either nonparametric or permutation tests. Use those tests instead.
  - Don't remove data in order to use approximate tests developed in the 1930-40s. Use more modern methods.

© 2020 PracticalStats.com

16

16

Applied Environmental Statistics Top Twelve Tips

2

Treat outliers like children …

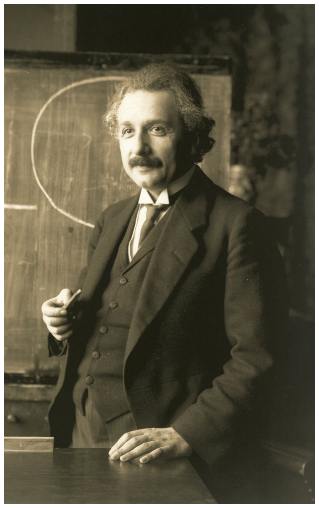
correct them when needed but never throw them out.

(see http://www.practicalstats.com/top12tips/)

© 2020 PracticalStats.com

Summary for Outlier VID

- Outlier tests cannot tell you whether data are 'wrong', only that they aren't likely to have come from a normal distribution
- Environmental field data (water, air, soils, rock, biota) are usually skewed distributions, not following a normal distribution. There are physical reasons for this. Outliers are to be expected.
- Don't just delete outliers.
  Albert Einstein was an outlier.



© 2020 PracticalStats.com