The end of statistical significance?

Eric P. Smith

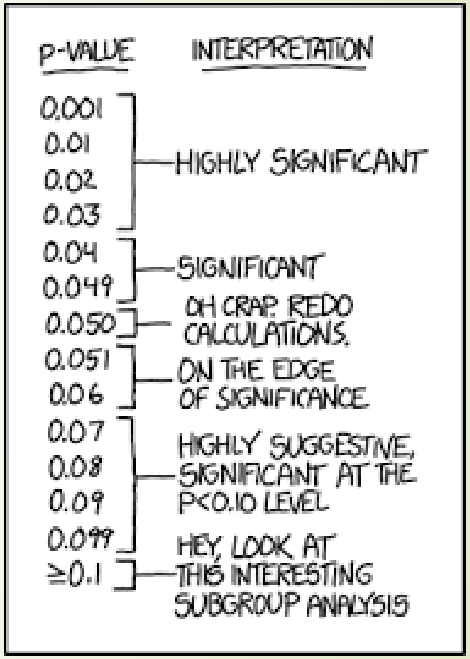
Statistics Department

Virginia Tech

CERF 2019, Mobile, Alabama, 7 November 2019

Believe it or not

- Statisticians have been arguing about (ie discussing) the value of statistical significance and p-values for years
  - Statistical significance creates a binary decision
    - p=0.051 = do not reject; p=0.049 = reject null hypothesis
  - p=0.05 is arbitrary and was based on creating tables of critical values
  - With wider data sets (ie more variables) it is easy to search for something “statistically significant” – p hacking
  - Statistical significance is often misunderstood and misinterpreted
    - IT IS NOT the probability the null is true



P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

https://www.ibm.com/developerworks/community/blogs/jfp/entry/Green_dice_are_loaded_welcome_to_p_hacking?lang=en

What to do? ASA Recommendations (Wasserstein and Lazar 2016)

1. P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

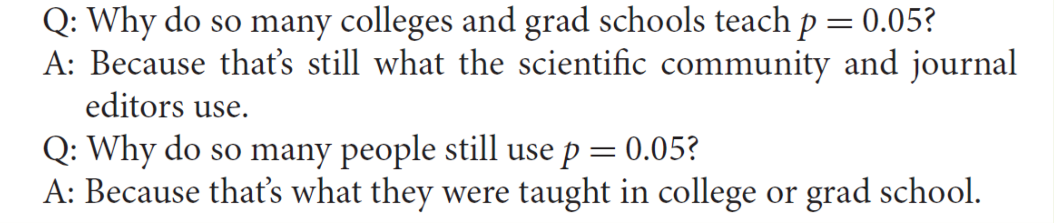
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

2019 40+ papers with suggestions

Maybe we should change however, we love statistical significance

- Relatively easy process given computing,
- Safety net
- Helps get us published
- Means we don't have to think about biological significance or effect size (just report significance)



Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

George Cobb

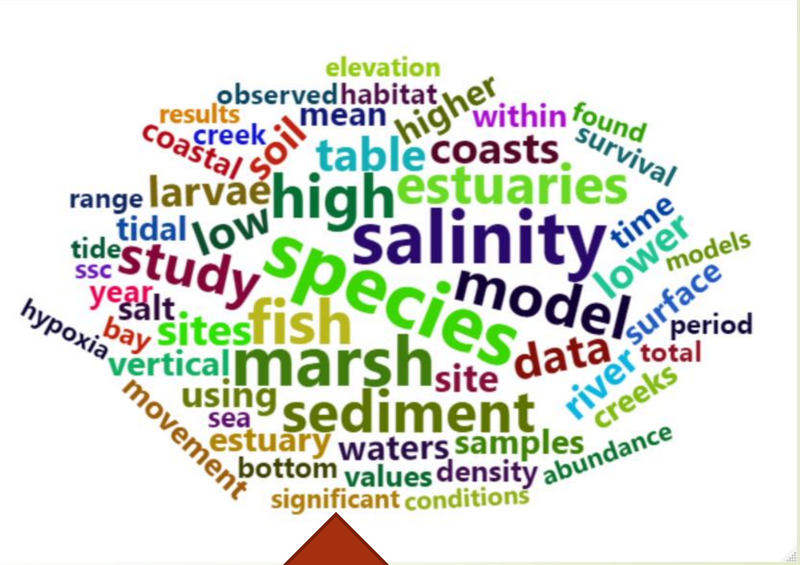
The significance addiction: Are papers in *Coasts and Estuaries* different?

How often is "significant" used?

19 papers published in 2019[1]

Significant/significance or significantly used 238 times

roughly 12-13 times per paper



significance	18
significant	156
significantly	64

[1] accessed August 20, 2019 articles from 42:1419-1557

Is there evidence of carelessness/variability – some

The isolated p-value

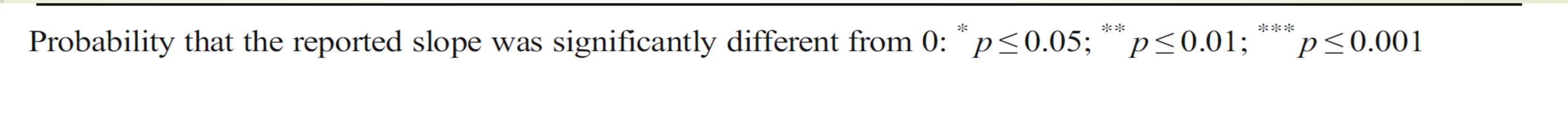
“The Shapiro-Wilk test revealed that assumption of normality was not achieved for salinity ($p < 0.05$), river discharge ($p < 0.05$), and chlorophyll a concentration ($p <0.05$).”

Is it a confidence interval or a test

“Pearson Correlation Matrix of correlation of water parameters and biological data during 2016 (correlations that are significant at the 95% level are shown in boldface)” (emphasis added, should be-5%)

Example

- Misinterpretation of p-value



Probability that the reported slope was significantly different from 0: ${}^{*}p \leq 0.05$; ${}^{**}p \leq 0.01$; ${}^{***}p \leq 0.001$

Example

- Significance is not always strong evidence.
- Large sample size results in significance although only 5% of variance is explained.
- Does the p-value tell you anything in this case?

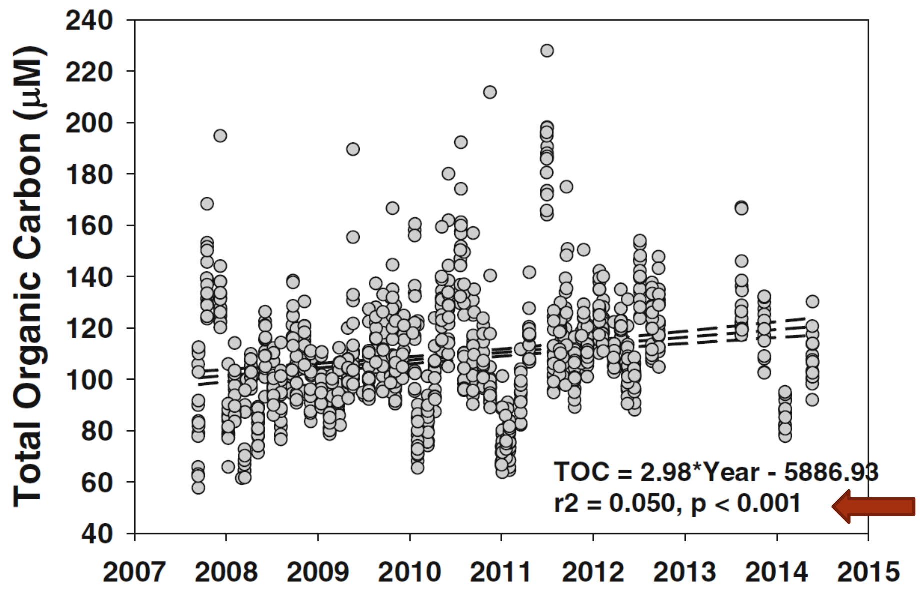


Fig. 2 Time series of TOC collected at all 17 sites during the period 2007–2014. Linear regression plotted, dashed lines indicate 95% confidence interval

Examples

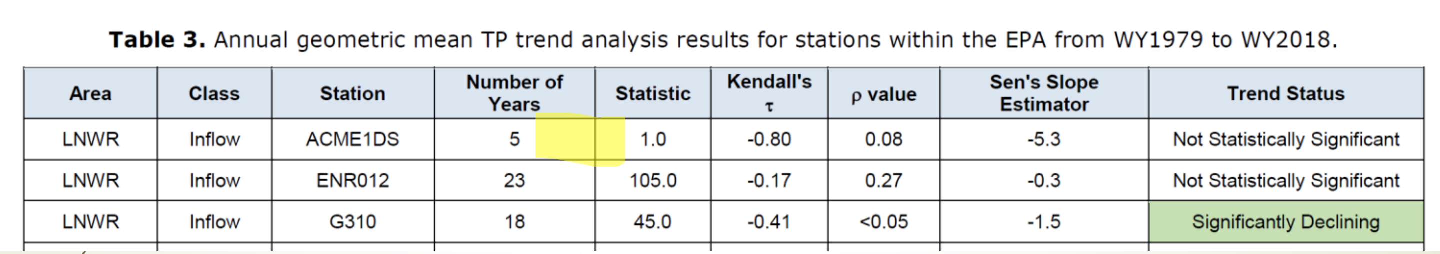


Table 3. Annual geometric mean TP trend analysis results for stations within the EPA from WY1979 to WY2018.

Area	Class	Station	Number of Years	Statistic	Kendall's τ	ρ value	Sen's Slope Estimator	Trend Status
LNWR	Inflow	ACME1DS	5	1.0	-0.80	0.08	-5.3	Not Statistically Significant
LNWR	Inflow	ENR012	23	105.0	-0.17	0.27	-0.3	Not Statistically Significant
LNWR	Inflow	G310	18	45.0	-0.41	<0.05	-1.5	Significantly Declining

From SFER 2019 report

Is this needed?
Note size of tau
for first row

Examples

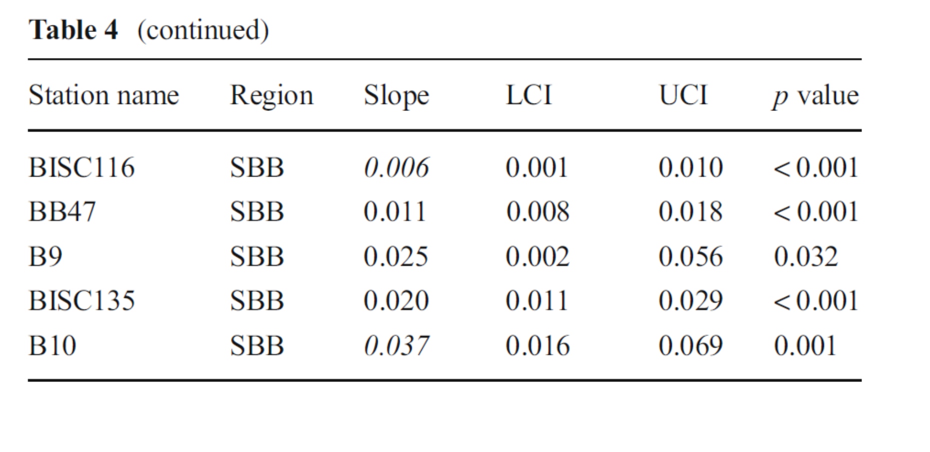


Table 4 (continued)

Station name	Region	Slope	LCI	UCI	*p* value
BISC116	SBB	*0.006*	0.001	0.010	< 0.001
BB47	SBB	0.011	0.008	0.018	< 0.001
B9	SBB	0.025	0.002	0.056	0.032
BISC135	SBB	0.020	0.011	0.029	< 0.001
B10	SBB	*0.037*	0.016	0.069	0.001

"We report all of the slopes along with their 95% confidence intervals and p values in place of labeling a slope as significant or not significant. This was done because waiting for a slope to be considered significant based on an arbitrary criterion can increase management response time to a system that is likely experiencing significant shifts in water quality (e.g., the Precautionary Principle, Raffensperger and Tickner 1999)."

Millette et al., 2019

So what to do: Reporting options

- Report p-value and provide evidence to support your decision/results
- Use significance in designed experiments not observational studies
- Report confidence intervals
- Focus on estimation/modeling rather than testing
- Report effect sizes: For a calculator see: https://www.psychometrica.de/effect_size.html
- R package **effsize, compute.es, sjstats, lsr, pwr**
- New graphical displays of data (see Ho et al., 2019, DABEST)

OR just go Bayesian calculate P(H|data) not P(data|H)

Reporting – add details to a supplemental

“The Shapiro-Wilk test revealed that assumption of normality was not achieved for salinity ($p < 0.05$), river discharge ($p < 0.05$), and chlorophyll a concentration ($p <0.05$).”

Normality was evaluated with graphical methods and the Shaprio-Wilk test on the RESIDUALS and we found the need to transform salinity, river discharge and Chla. Details are in the supplemental material.

Reporting – avoid just reporting p-value

“From May 2016 to December 2016, overall conductivity during high tide has decreased and has significantly changed compared with sampling during 2011 (p value < 0.05, Fig. 3).”

“Overall conductivity changed with tide and year combinations (KW=4.6, p value =0.003, n=?, Fig. 3). Boxplots in Fig 3. Illustrate the … ”

“Plant height decreased linearly with elevation for S. patens (r2 = .305, p < .05),”

Plant height decreased linearly with elevation for S. patens (r^2 = .305, slope = 14.2, 95%CI 11.4 to 17, n=?),

Reporting

"ANOVA analysis for M4 revealed that only month (p = 0.02) and year (p = 0.04) had significant effects on the residual condition index."

"ANOVA analysis for M4 revealed that only month ($F_{10,41}$=3.6, p = 0.02) and year

($F_{4,41}$=2.7, p = 0.04) had mild effects on the residual condition index."

Note one could also use an AIC approach here.

Reporting

“Pearson Correlation Matrix of correlation of water parameters and biological data during 2016 (correlations that are significant at the **95%** level are shown in boldface)” (emphasis added, should be 5%)

“Pearson Correlation Matrix of water parameters and biological data during 2016 (correlations that are greater than 0.5 are in boldface). P-values are given below correlations along with sample sizes.

Example: reporting a result that does not pass the 0.05

"The 6-month treatment difference, using ANCOVA to take into account baseline office systolic BP, was 4.11 mm Hg (95% CI: 8.44 to 0.22; p = 0.064) (Table 1), similar to the unadjusted difference, but with an anticipated slight increase in precision (i.e., the CI is smaller and the p value is lower)."

Pocock et al 2016 J Am Col Cardiology, 2016-25

Option - report effect size: a measure of the magnitude of the phenomenon

Effect size	r correlation coefficient
Small	0.10
Medium	0.30
Large	0.50

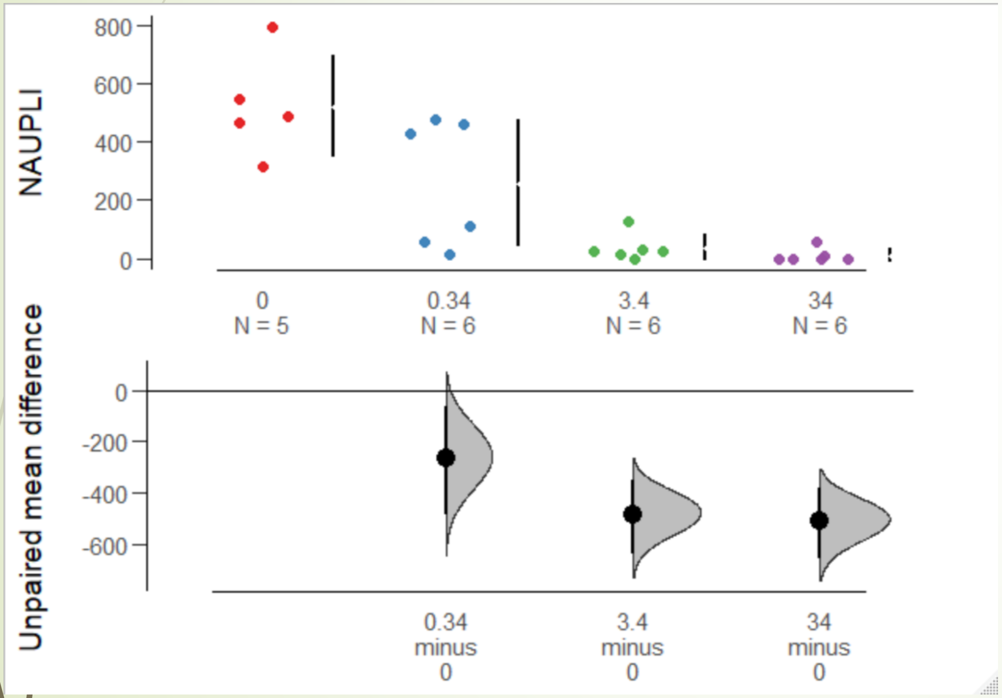
$$d^2 = r^2/(1-r^2)$$

Effect size – mean difference	d
Very small	0.01
Small	0.20
Medium	0.50
Large	0.80
Very large	1.20
Huge	2.0

$$d = \frac{\overline{x_1} - \overline{x_2}}{s}$$

Ex: Condition index was higher in restored site than in the reference site: mean difference =1.2 units (95% interval 0.6 to 1.8, p=0.0002, effect size=0.53)

Option: graphical displays Mesocosm study: 0 is the control



```
DABEST (Data Analysis with Bootstrap Estimation) v0.2.2
=======================================================

Variable: NAUPLI

Unpaired mean difference of 0.34 (n=5) minus 0 (n=6)
 -263 [95CI  -488; -61.1]

Unpaired mean difference of 3.4 (n=5) minus 0 (n=6)
 -483 [95CI  -637; -353]

Unpaired mean difference of 34 (n=5) minus 0 (n=6)
 -509 [95CI  -659; -378]

5000 bootstrap resamples.
All confidence intervals are bias-corrected and accelera
ted
```

Graphic from dabest package in R

Regression options – mesocosm data

A regression analysis using dose as the explanatory variable resulted in a linear regression with intercept 275.4 and slope -8.4 units (95% interval)

The regression summaries

- CI: -14.76 -2.03
- Standardized coefficient: -0.513
- d estimate: 1.07 (large)
- A simple linear regression resulted in an estimated model Nauplii = 275 – 8.4*dose (95% CI for slope -14.76 to -2.03, n=23)
- Regression analysis suggested a strong effect of dose on Nauplii (slope = -8.4, 95% CI -14.76 to -2.03, effect size = 1.07, n=23)

Issues with change in policy?

- Effect sizes for some tests may be not clear: nonparametric tests (seasonal Kendall), normality checks, Generalized additive models
- Should we use multiple comparison methods?
- Should we use power analysis for sample size calculations?
- Should we adjust p-values for multiple testing?
- Multivariate tests/normality tests/other tests
- How to report Bayesian analysis?

Other issues

- Transparency
- Confirmation bias
- Reproducibility and replicability
  - (see Beck et al 2019 Estuaries and Coasts 42:1774-1791).
- Correlated observations
- "Found data"

References

- Millette, N.C., et al. Using Spatial Variability in the Rate of Change of Chlorophyll a to Improve Water Quality Management in a Subtropical Oligotrophic Estuary. *Estuaries and Coasts* (2019). DOI: 10.1007/s12237-019-00610-5
- Wasserstein R.L. & Lazar, N.A. (2016). The ASA Statement on *p*-Values: Context, Process, and Purpose, *The American Statistician* 70(2):129-133 DOI: 10.1080/00031305.2016.1154108
- Wasserstein, R.L., Schirm, A.L. & Lazar, N.A. (2019) Moving to a World Beyond "$p < 0.05$", *The American Statistician* 73(Sup1): 1-19. DOI: 10.1080/00031305.2019.1583913
- Kyriacou, D.M. (2016). The Enduring Evolution of the *P* Value *JAMA*. 315(11):1113-1115. DOI: 10.1001/jama.2016.2152
- Lenhard, W. & Lenhard, A. (2016). Calculation of Effect Sizes. *Psychometrica*. Dettelbach (Germany). DOI: 10.13140/RG.2.1.3478.4245 Retrieved from: https://www.psychometrica.de/effect_size.html
- Ho, J., et al. (2019). Moving beyond *P* values: data analysis with estimation graphics. *Nature Methods* 16:565–566 DOI: 10.1038/s41592-019-0470-3
- Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1): 7–29. https://doi.org/10.1177/0956797613504966
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Guidance: Psychological Science

https://www.psychologicalscience.org/publications/psychological_science/ps-submissions

- **Statistics**
  Psychological Science recommends the use of the "new statistics"—effect sizes, confidence intervals, and meta-analysis—to avoid problems associated with null-hypothesis significance testing (NHST). Authors are encouraged to consult this *Psychological Science* tutorial by Geoff Cumming, which argues that estimation and meta-analysis are more informative than NHST and that they foster development of a cumulative, quantitative discipline. Cumming has also prepared a video workshop on the new statistics that can be found here.
- Authors must include effect sizes for their major results and distributional information in their graphs (or tables, for that matter). Fine-grained graphical presentations that show how data are distributed are often the most honest way of communicating results. Please report 95% confidence intervals instead of standard deviations or standard errors around mean dependent variables, because confidence intervals convey more useful information—another point discussed in Cumming's tutorial.
- **Reporting Statistical Results**
- The abstract should include information about the sample size(s) in studies reported in the manuscript. Please report test statistics with two decimal points (e.g., t(34) = 5.67) and probability values with three decimal points. In addition, exact p values should be reported for all results greater than .001; p values below this range should be described as "p < .001." Authors should be particularly attentive to APA style when typing statistical details (e.g., *N*s for chi-square tests, formatting of *df*s), and if special mathematical expressions are required, they should not be graphic objects but rather inserted with Word's Equation Editor or similar.

Guidance: NEJM

https://www.nejm.org/author-center/new-manuscripts

- Our Statistical Consultants recommend the following best statistical practices in manuscripts submitted to the *Journal*. We recommend that you follow them in the design and reporting of research studies.
- **For all studies:**
- The Methods section of all manuscripts should contain a brief description of sample size and power considerations for the study, as well as a brief description of the methods for primary and secondary analyses.
- The Methods section of all manuscripts should include a description of how missing data have been handled. Unless missingness is rare, a complete case analysis is generally not acceptable as the primary analysis and should be replaced by methods that are appropriate, given the missingness mechanism. Multiple imputation or inverse probability case weights can be used when data are missing at random; model-based methods may be more appropriate when missingness may be informative. For the *Journal's* general approach to the handling of missing data in clinical trials please see Ware et al (N Engl J Med 2012;367:1353–1354).
- Significance tests should be accompanied by confidence intervals for estimated effect sizes, measures of association, or other parameters of interest. The confidence intervals should be adjusted to match any adjustment made to significance levels in the corresponding test.
- Unless one-sided tests are required by study design, such as in noninferiority clinical trials, all reported P values should be two-sided. In general, P values larger than 0.01 should be reported to two decimal places, and those between 0.01 and 0.001 to three decimal places; P values smaller than 0.001 should be reported as P<0.001. Notable exceptions to this policy include P values arising from tests associated with stopping rules in clinical trials or from genome-wide association studies.
- There's more ….

Other examples of isolated p-values

“From May 2016 to December 2016, overall conductivity during high tide has decreased and has significantly changed compared with sampling during 2011 (p value < 0.05, Fig. 3).”

“Subsidence in S. alterniflora pots was significantly lower than in unplanted controls (p < .01; Fig. 8).”

“ANOVA analysis for M4 revealed that only month (p = 0.02) and year (p = 0.04) had significant effects on the residual condition index.”