Statistics Pitfalls and Tips

Darrell Sonntag

CE Graduate Seminar

October 17, 2024

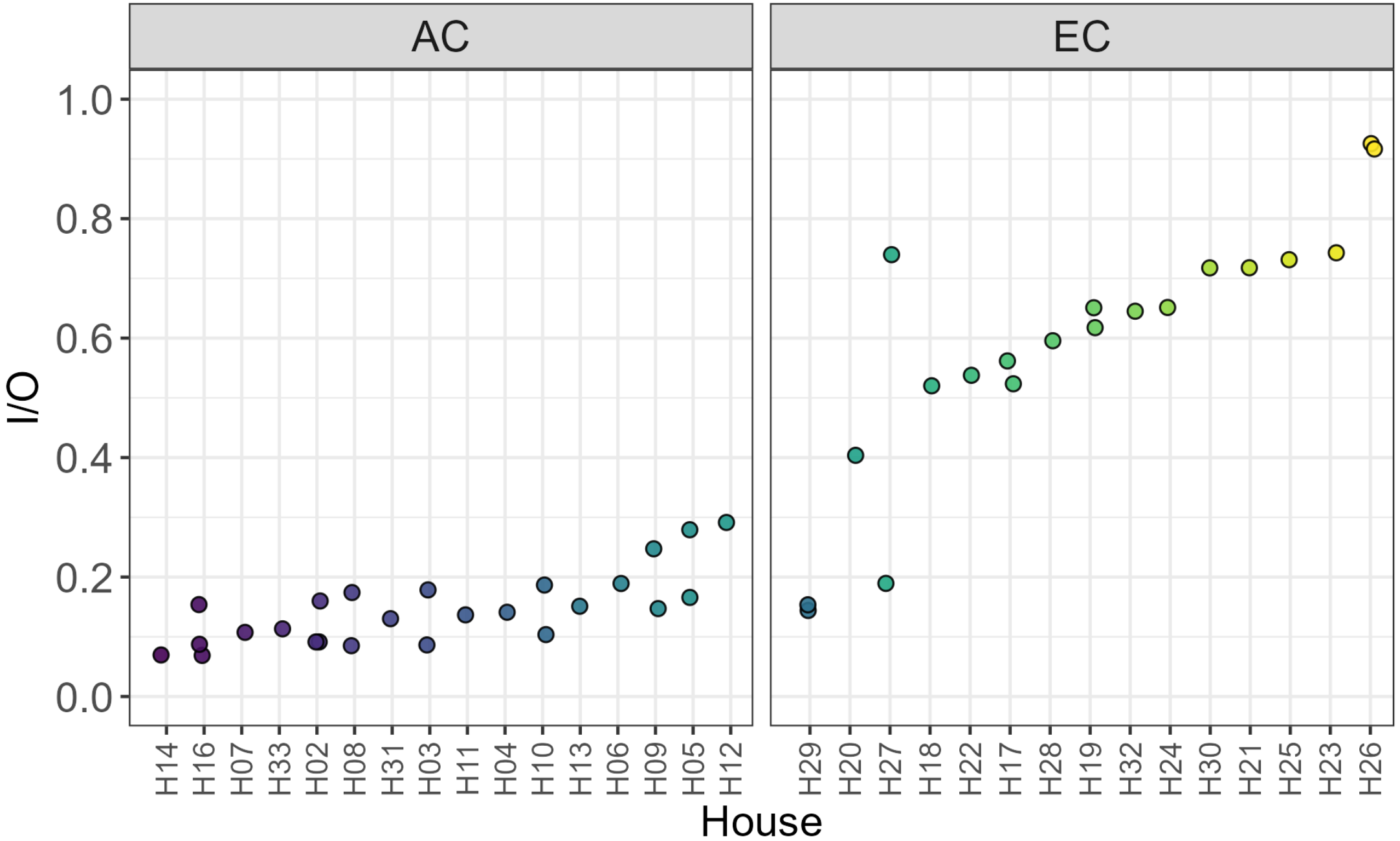
Why are do we care about statistical tests?

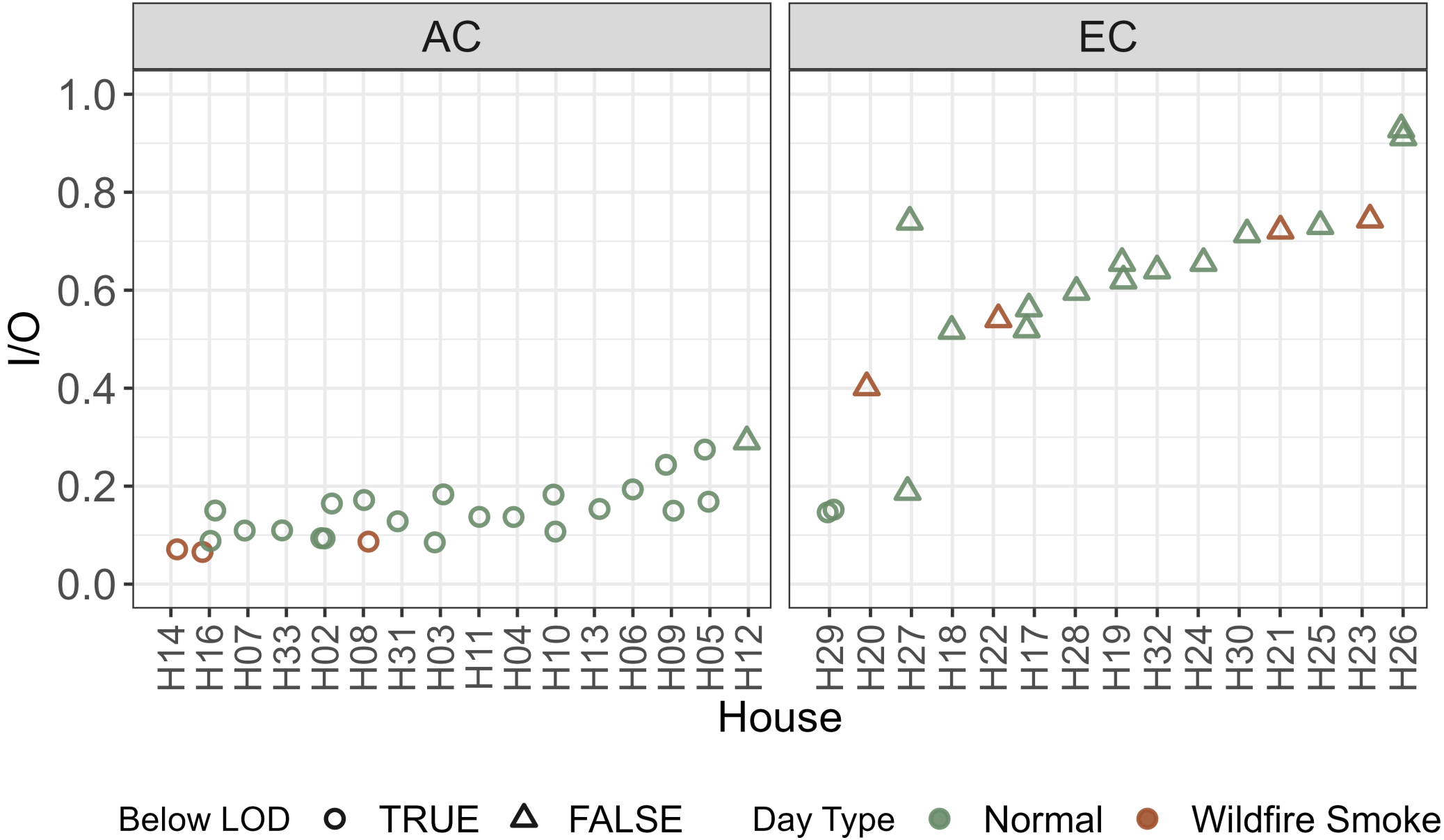
Why are do we care about statistical tests?

- Are the differences we see in our data due to random variability of our sample?
- Or do the data suggest that the differences are due to true differences between groups?
  - Strength of different treatments of concrete
  - Air quality in different types of homes
  - etc.

Question #1

- Download ozone.IO.csv
- Calculate the mean indoor/outdoor (I/O) ratio of ozone concentrations for Central and Evaporative Air-conditioned homes
- Calculate the 95% confidence intervals of the mean for each type of home


AC
EC
1.0
0.8
0.6
0.4
0.2
0.0
I/O
H14
H16
H07
H33
H02
H08
H31
H03
H11
H04
H10
H13
H06
H09
H05
H12
H29
H20
H27
H18
H22
H17
H28
H19
H32
H24
H30
H21
H25
H23
H26
House


AC
EC
I/O
1.0
0.8
0.6
0.4
0.2
0.0
H14
H16
H07
H33
H02
H08
H31
H03
H11
H04
H10
H13
H06
H09
H05
H12
H29
H20
H27
H18
H22
H17
H28
H19
H32
H24
H30
H21
H25
H23
H26
House
Below LOD
TRUE
FALSE
Day Type
Normal
Wildfire Smoke

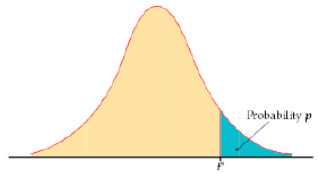
Confidence interval using t-distribution

$$\left(\bar{x} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$$

- $\bar{x}$ = sample mean
- s = sample standard deviation
- n = sample size
- $t_{n-1,\alpha/2}$ = t critical value (next slide)
- $\propto$ = type I error (typically 5%)
- $1 - \propto$ = Confidence level (typically 95%)

t-distribution table

Areas in the upper tail are given along the top of the table. Critical t* values are given in the table.



df	0.1	0.05	0.025	0.02	0.01	0.005
1	3.078	6.314	12.706	15.895	31.821	63.657
2	1.886	2.920	4.303	4.849	6.965	9.925
3	1.638	2.353	3.182	3.482	4.541	5.841
4	1.533	2.132	2.776	2.999	3.747	4.604
5	1.476	2.015	2.571	2.757	3.365	4.032
6	1.440	1.943	2.447	2.612	3.143	3.707
7	1.415	1.895	2.365	2.517	2.998	3.499
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
10	1.372	1.812	2.228	2.359	2.764	3.169
11	1.363	1.796	2.201	2.328	2.718	3.106
12	1.356	1.782	2.179	2.303	2.681	3.055
13	1.350	1.771	2.160	2.282	2.650	3.012
14	1.345	1.761	2.145	2.264	2.624	2.977
15	1.341	1.753	2.131	2.249	2.602	2.947
16	1.337	1.746	2.120	2.235	2.583	2.921
17	1.333	1.740	2.110	2.224	2.567	2.898
18	1.330	1.734	2.101	2.214	2.552	2.878
19	1.328	1.729	2.093	2.205	2.539	2.861
20	1.325	1.725	2.086	2.197	2.528	2.845
21	1.323	1.721	2.080	2.189	2.518	2.831
22	1.321	1.717	2.074	2.183	2.508	2.819
23	1.319	1.714	2.069	2.177	2.500	2.807
24	1.318	1.711	2.064	2.172	2.492	2.797
25	1.316	1.708	2.060	2.167	2.485	2.787
26	1.315	1.706	2.056	2.162	2.479	2.779
27	1.314	1.703	2.052	2.158	2.473	2.771
28	1.313	1.701	2.048	2.154	2.467	2.763
29	1.311	1.699	2.045	2.150	2.462	2.756
30	1.310	1.697	2.042	2.147	2.457	2.750
31	1.309	1.696	2.040	2.144	2.453	2.744
32	1.309	1.694	2.037	2.141	2.449	2.738
33	1.308	1.692	2.035	2.138	2.445	2.733
34	1.307	1.691	2.032	2.136	2.441	2.728
35	1.306	1.690	2.030	2.133	2.438	2.724
36	1.306	1.688	2.028	2.131	2.434	2.719
37	1.305	1.687	2.026	2.129	2.431	2.715
38	1.304	1.686	2.024	2.127	2.429	2.712
39	1.304	1.685	2.023	2.125	2.426	2.708
40	1.303	1.684	2.021	2.123	2.423	2.704
41	1.303	1.683	2.020	2.121	2.421	2.701
42	1.302	1.682	2.018	2.120	2.418	2.698
43	1.302	1.681	2.017	2.118	2.416	2.695
44	1.301	1.680	2.015	2.116	2.414	2.692
45	1.301	1.679	2.014	2.115	2.412	2.690
46	1.300	1.679	2.013	2.114	2.410	2.687
47	1.300	1.678	2.012	2.112	2.408	2.685
48	1.299	1.677	2.011	2.111	2.407	2.682
49	1.299	1.677	2.010	2.110	2.405	2.680
50	1.299	1.676	2.009	2.109	2.403	2.678

df	0.1	0.05	0.025	0.02	0.01	0.005
51	1.298	1.675	2.008	2.108	2.402	2.676
52	1.298	1.675	2.007	2.107	2.400	2.674
53	1.298	1.674	2.006	2.106	2.399	2.672
54	1.297	1.674	2.005	2.105	2.397	2.670
55	1.297	1.673	2.004	2.104	2.396	2.668
56	1.297	1.673	2.003	2.103	2.395	2.667
57	1.297	1.672	2.002	2.102	2.394	2.665
58	1.296	1.672	2.002	2.101	2.392	2.663
59	1.296	1.671	2.001	2.100	2.391	2.662
60	1.296	1.671	2.000	2.099	2.390	2.660
61	1.296	1.670	2.000	2.099	2.389	2.659
62	1.295	1.670	1.999	2.098	2.388	2.657
63	1.295	1.669	1.998	2.097	2.387	2.656
64	1.295	1.669	1.998	2.096	2.386	2.655
65	1.295	1.669	1.997	2.096	2.385	2.654
66	1.295	1.668	1.997	2.095	2.384	2.652
67	1.294	1.668	1.996	2.095	2.383	2.651
68	1.294	1.668	1.995	2.094	2.382	2.650
69	1.294	1.667	1.995	2.093	2.382	2.649
70	1.294	1.667	1.994	2.093	2.381	2.648
71	1.294	1.667	1.994	2.092	2.380	2.647
72	1.293	1.666	1.993	2.092	2.379	2.646
73	1.293	1.666	1.993	2.091	2.379	2.645
74	1.293	1.666	1.993	2.091	2.378	2.644
75	1.293	1.665	1.992	2.090	2.377	2.643
76	1.293	1.665	1.992	2.090	2.376	2.642
77	1.293	1.665	1.991	2.089	2.376	2.641
78	1.292	1.665	1.991	2.089	2.375	2.640
79	1.292	1.664	1.990	2.088	2.374	2.640
80	1.292	1.664	1.990	2.088	2.374	2.639
81	1.292	1.664	1.990	2.087	2.373	2.638
82	1.292	1.664	1.989	2.087	2.373	2.637
83	1.292	1.663	1.989	2.087	2.372	2.636
84	1.292	1.663	1.989	2.086	2.372	2.636
85	1.292	1.663	1.988	2.086	2.371	2.635
86	1.291	1.663	1.988	2.085	2.370	2.634
87	1.291	1.663	1.988	2.085	2.370	2.634
88	1.291	1.662	1.987	2.085	2.369	2.633
89	1.291	1.662	1.987	2.084	2.369	2.632
90	1.291	1.662	1.987	2.084	2.368	2.632
91	1.291	1.662	1.986	2.084	2.368	2.631
92	1.291	1.662	1.986	2.083	2.368	2.630
93	1.291	1.661	1.986	2.083	2.367	2.630
94	1.291	1.661	1.986	2.083	2.367	2.629
95	1.291	1.661	1.985	2.082	2.366	2.629
96	1.290	1.661	1.985	2.082	2.366	2.628
97	1.290	1.661	1.985	2.082	2.365	2.627
98	1.290	1.661	1.984	2.081	2.365	2.627
99	1.290	1.660	1.984	2.081	2.365	2.626
100	1.290	1.660	1.984	2.081	2.364	2.626

Confidence interval using t-distribution

$$\left(\bar{x} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} \right)$$

- Assumptions:
  - Independent, random sample from normal population
  - “Robust to small or even moderate departures from normality unless n is quite small”

Confidence interval using t-distribution

$$\left(\bar{x} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$$

- Assumptions:
  - Independent, random sample from normal population
  - “Robust to small or even moderate departures from normality unless n is quite small”
- Is it appropriate for us to apply the t-distribution to our data?

Confidence interval using t-distribution

$$\left(\bar{x} - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$$

- How do we decrease the size of the confidence intervals?
  - Increase n (which also decreases $t_{n-1,\alpha/2}$)
  - Decrease standard deviation (s)

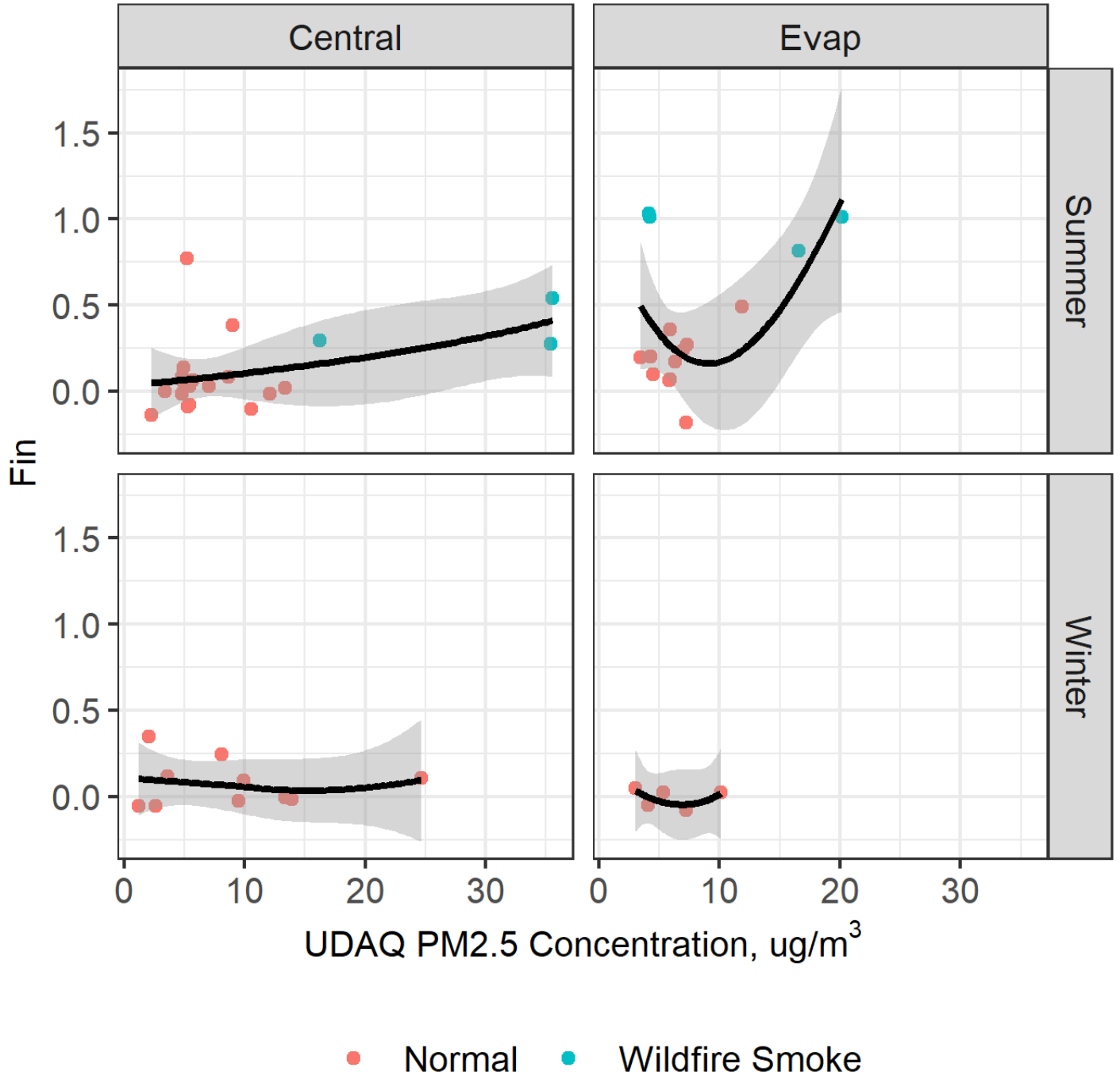
Question #2

- Download timeseries.csv
- Calculate the mean indoor aerosol concentrations for Central and Evaporative Air-conditioned homes
- Calculate the 95% confidence intervals of the mean for each type of home

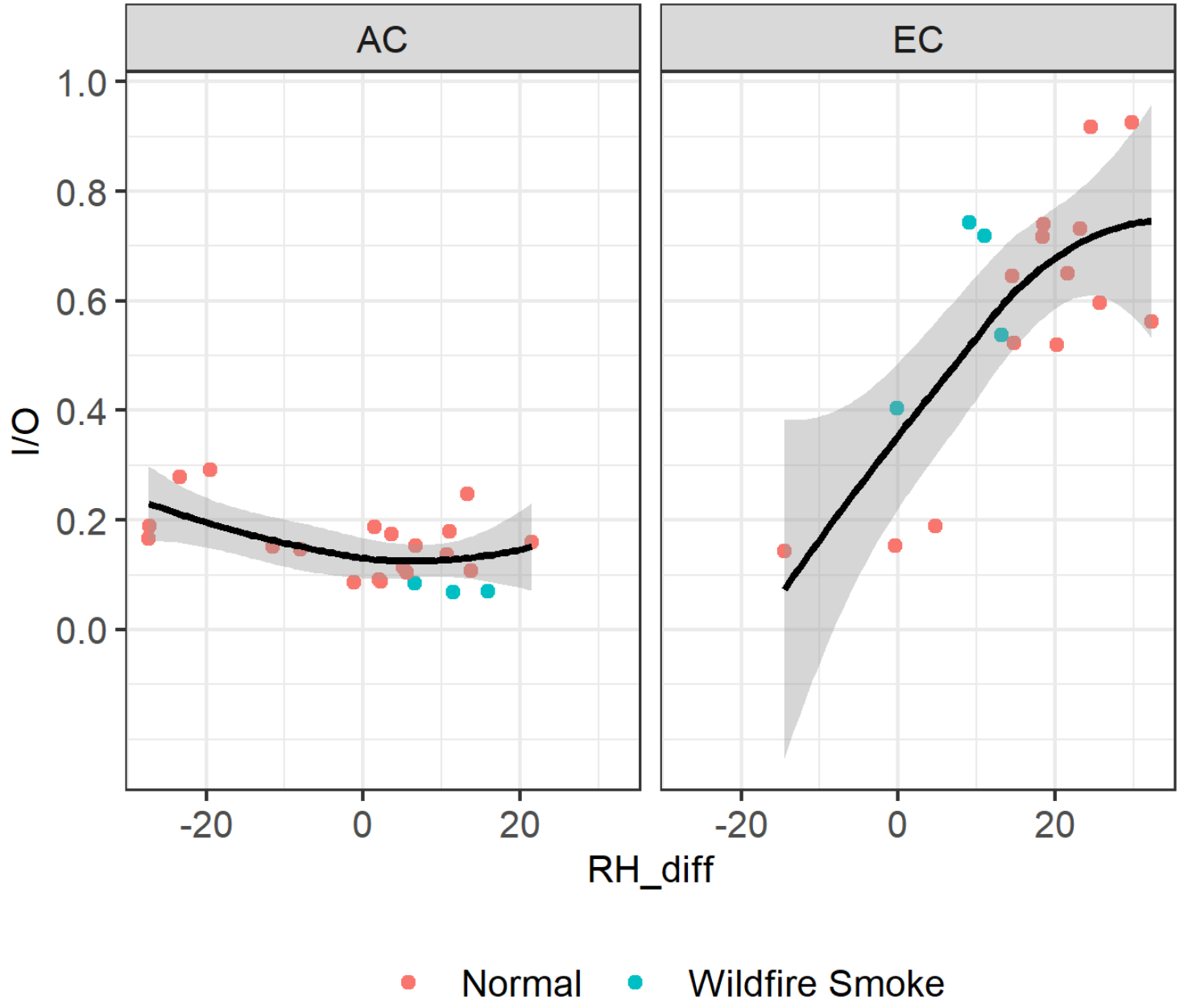
Tip 1: Graph your data (inside and out)

- Use statistical models and tests to confirm what you already suspect or know from graphs
- If you are learning about important relationships from your statistical tests, I have two concerns:
  - You probably haven’t sufficiently visualized your data
  - Are you sure your statistical assumptions of your results are satisfied?

Example: Wildfire Smoke



Example: Wildfire Smoke



Tip 2: Are the statistical tests appropriate for your random variable?

- What is a random variable?
  - A variable that can change with each observation
  - For example..
    - Traffic flow at same location at same time during the week
    - Strain measured on beam during a test
- Are my random variables independent?
  - Is my sampling truly random?
  - Are my observations dependent on one another? Or are they influenced by one? Or more likely to be similar or different than others?
- Are my random variables and identically distributed?
  - Do my observations come from the same probability distribution (e.g. normal)

Tip 3: Change your random variable (if needed)

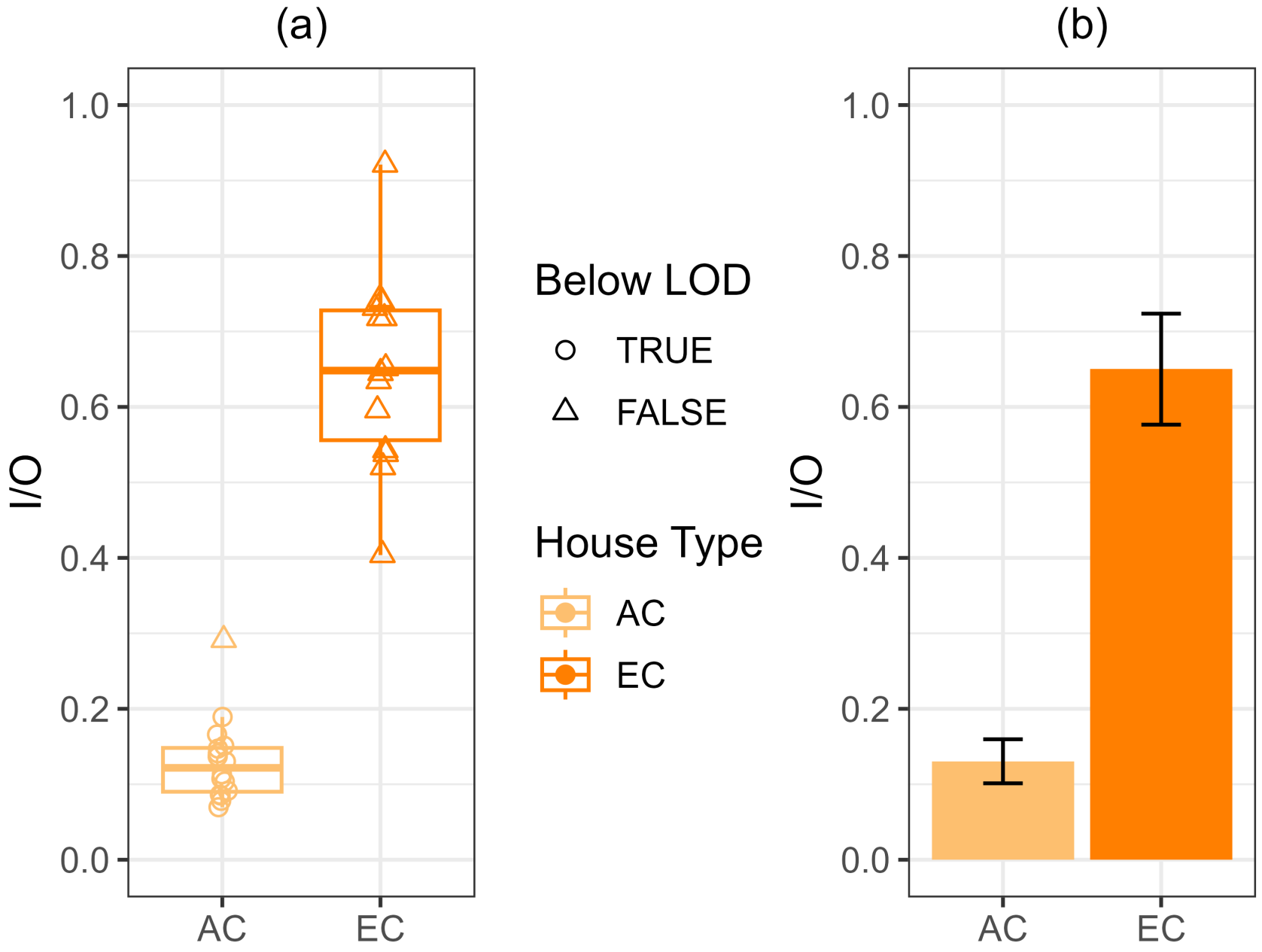
- If your random variable doesn’t meet the assumptions, you have options
  - Easier: Change your random variable of interest if your current random variables doesn’t meet the
  - Harder: Develop more complex statistical model (e.g. mixed model that accounts for dependence of samples and other factors, time-series model that accounts for autocorrelation)
- We went with the easier option above

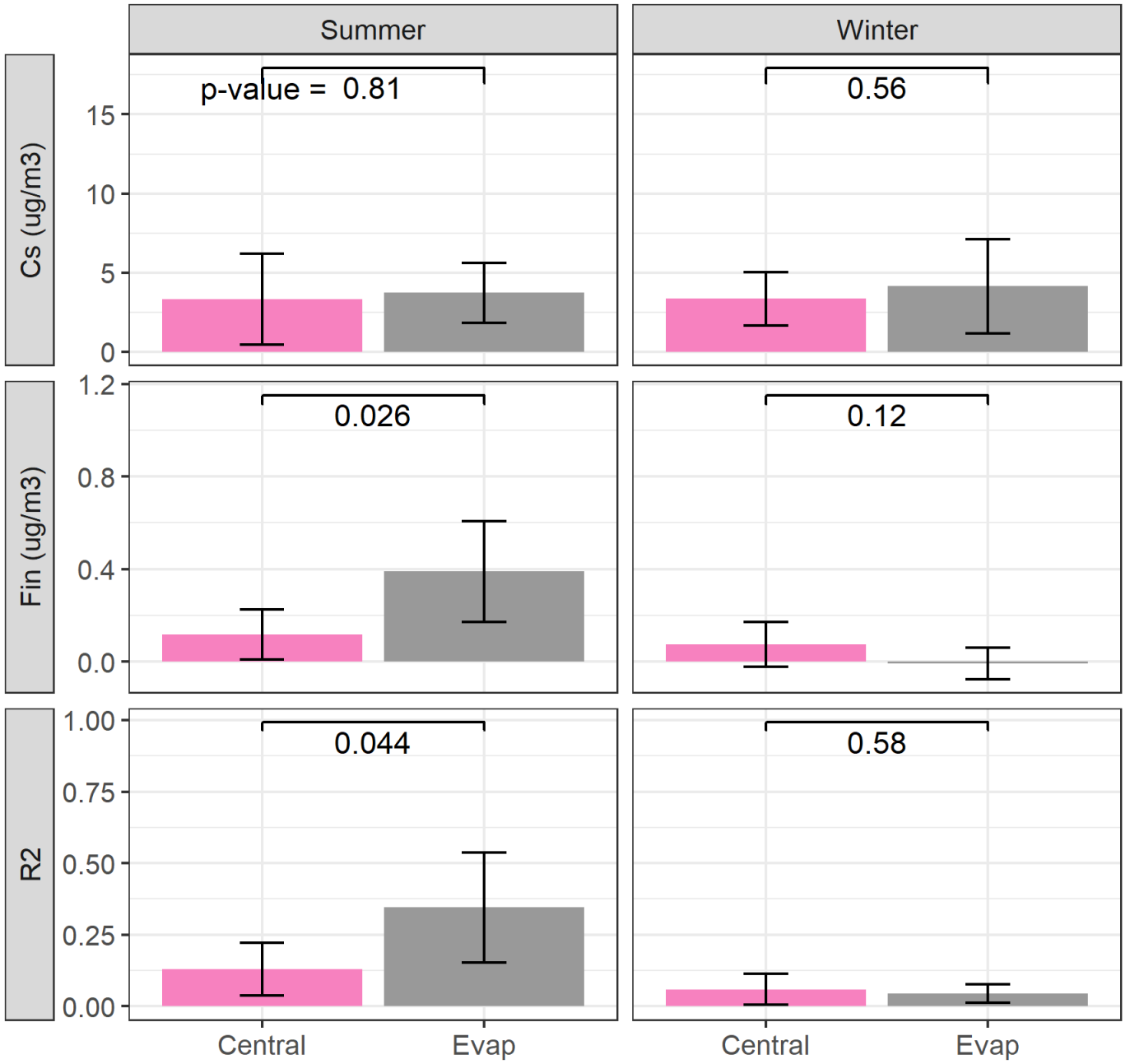
Tip 3: Change your random variable (if needed)

- Question #1
  - Initially, we chose to evaluate the average I/O from each visit as our random variable of interest
  - Next, we chose to evaluate the average I/O from each home (potentially averaged over several visits) as our random variable of interest
    - Now our random variables are independent from one another
- Question #2
  - Initially, we chose to evaluate the average 1-minute indoor concentrations from our data
    - Data from one visit were highly dependent on each other
  - Next, we averaged the 1-minute indoor concentrations for each visit, then the average visits from each house. We treated the average indoor concentrations at each house as our random variable of interest

Tip 4: Don’t rely on 95% Confidence intervals to determine statistical significance

- IF the 95% CI’s overlap, does it mean the data are not significantly different at a 5% type I error rate?
- OR what if one group has a large confidence interval, but not the other group?
  - You don’t know for sure….
- Well, if they don’t overlap at all, I know they are significantly different
  - But…this is overly conservative, and there may be significant differences
- Solution:
  - Better to use a t-test
  - Show results of t-test graphically


(a)
(b)
I/O
1.0
0.8
0.6
0.4
0.2
0.0
AC
EC
Below LOD
TRUE
FALSE
House Type
AC
EC


Summer
Winter
Cs (ug/m3)
Fin (ug/m3)
R2
p-value = 0.81
0.56
0.026
0.12
0.044
0.58
15
10
5
0
1.2
0.8
0.4
0.0
1.00
0.75
0.50
0.25
0.00
Central
Evap
Central
Evap

Tip 5: Use scripts

- Keep your graphics and analysis separate from your data files
- Advantages:
  - You have a record of your analysis steps from raw data to final results for paper
  - You can easily apply your analysis to new datasets

Tip 6: Take my class

- CE 594R Winter 2025
- Data Science for Engineers
- Objectives:
  - Learn to script using R
  - Large focus on data visualization (graphics)
  - Apply statistical models
  - Learn machine learning techniques
  - Bring your own research data for the course project