Discussion 13: Student Questions & Peer Review

Arpitha / Coulton / Haley

Statistics: Spatial Data

- If your data has different locations (e.g. measurements at Barton Pond and Allen Creek), you may want to treat the location as **categories**
  - You can analyze these using ANOVA, a t-test, etc.

Statistics: T-Test

- A t-test is a method to determine if the mean of two samples are equal
  - Assumes that the two samples have **equal variances**
  - If the variances aren’t equal use Welch’s paired t-test

Statistics: T-Test

- Advantage: we can say more about the relationship between the two samples than if we use ANOVA
  - We can test if sample 1 > sample 2, sample 1 < sample 2, or sample 1 ≠ sample 2
  - Running a regression with one variable is equivalent to a t-test where the null hypothesis is sample 1 ≠ sample 2
- Disadvantage: we can only test two samples at a time

Statistics: Two Sample T-Test

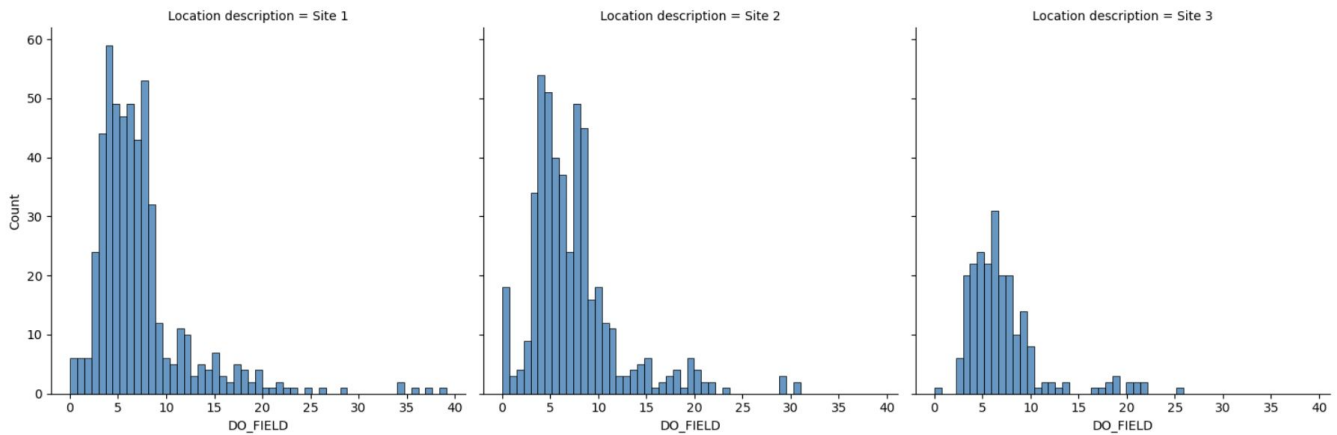
Example using Barton Pond monitoring data

Clean data & check for equal variances

[5]:
```
df = df[df['DO_FIELD'] < 100]
df['Location description'] = df['Location description'].apply(lambda s: np.where('Site 1' in s, 'Site 1',
                                                                                  np.where('Site 2' in s, 'Site 2', 'Site 3')))
```

[6]:
```
sns.displot(data = df, x = 'DO_FIELD', col = 'Location description')
```

[6]:
```
<seaborn.axisgrid.FacetGrid at 0x14d7f7810>
```



Statistics: Two Sample T-Test

Example using Barton Pond monitoring data

Run test

- Output is the t-test statistics, the p-value and degrees of freedom

[7]:
```
site_1 = df[df['Location description'] == 'Site 1']['DO_FIELD']
site_2 = df[df['Location description'] == 'Site 2']['DO_FIELD']

sm.stats.ttest_ind(site_1, site_2)
```

[7]:
```
(-0.37368955791325514, 0.7087153042740553, 990.0)
```

Statistics: Paired T-Test

We use this when our data points are pairs:

- e.g before and after measurements
- matched pair clinical studies

Implemented in scipy

```
from scipy.stats import ttest_rel

# Python paired sample t-test
ttest_rel(a, b)
```

Statistics: Dealing with null results

In our t-test example, the p-value was 0.71. This is not statistically significant.

This is okay! Focus on:

- Making sure your statistical tests are appropriate for your data
- Interpreting the results in context

Statistics: Correlations

There is no p-value for a correlation coefficient

To decide if a correlation is meaningful we typically:

- Use domain knowledge: a high correlation coefficient means different things in medicine vs psychology, because the types of relationships we observe and the strength of evidence we need is different
- Use heuristics: one common heuristic is that $r > 0.7$ is large
    - This is weaker than use domain knowledge
- Use another statistical test: for example, we may run a regression on the two variables

Peer Review