Python Statistical Testing Cheat Sheet

I want to compare two variables:

Var 1	Var 2	You Should Use
Continuous	Continuous	Pearson's R
Continuous	Categorical	T-Test
Categorical	Categorical	χ^2

---

p value $P(\text{data}|H_0)$

- If $p < \alpha$, we reject H_0
- If $p > \alpha$, we fail to reject H_0
- Keep in mind, a p value doesn't tell us anything about effect size

alpha, α 1 - our confidence level, typically .05, .01, or .001

Continuous v Categorical: T-Test

- `ttest_1samp(subgroup: List[float], pop_mean: float)`

  To compare a subgroup's mean against the population mean

  ```
  subgroup = df[df.some_group == 'a_value'].cont_var
  tstat, p = stats.ttest_1samp(subgroup, df.cont_var.mean())
  ```

- `ttest_ind(subgroup_1: List[float], subgroup_2: List[float])`

  To compare the means of two subgroups

  ```
  subgroup_1 = df[df.some_group == 'value one'].cont_var
  subgroup_2 = df[df.some_group == 'value two'].cont_var
  tstat, p = stats.ttest_ind(subgroup_1, subgroup_2)
  ```

T-Test

- Compares means for some continuous variable across 2 groups (or 1 group vs a population mean)
- Assumes the continuous variable is normally distributed; equal variance in both groups
- H_0: the means are the same

Pearson's R

- Measures how two continuous variables change together linearly
- Our test statistic, r, gives us the strength and direction of the **linear** correlation between the two vars
- H_0: there is no linear correlation between the two variables

---

Continuous v Continuous: Pearson's R

- `pearsonr(col_a: List[float], col_b: List[float])`

```
r, p = stats.pearsonr(df.col_a, df.col_b)
```

Categorical v Categorical: χ^2 Contingency

- `chi2_contingency(ctab: DataFrame)`

```
ctab = pd.crosstab(df.col_1, df.col_2)
chi2_stat, p, degf, expected = stats.chi2_contingency(ctab)
```

χ^2 Contingency

- Used to compare two categorical variables
- Compares observed vs expected proportions
- H_0: The two variables are independent