To determine if effects seen in the sample appear in the larger population.

- can use classical hypothesis testing
  there is also,
  - Fisher null hypothesis testing
  - Neyman-Pearson decision theory
  - Bayesian Inference

Classical Hypothesis Testing achieves,
given a sample and an apparent effect, what is the probability of seeing such an effect by chance.

① Quantify the size of the apparent effect by choosing a test statistic (e.g. a diff in means)

② Define a Null Hypothesis: a model of the real system based on the effect not being real (e.g. the distributions between both groups is identical)

③ Compute the p-value — the probability of seeing the apparent effect if the null hypothesis is true.

④ Interpret result If p-value is low,
the result is statistically significant —
the effect is more likely to appear in the larger population.

⇒ Similar to proof by contradiction.

Null Hypothesis Test. for Sample Data

Coin Toss Example — that is also sampled from a population which you are trying to make inference

Toss a coin 99 H=40, T=60.

Create Model 50:50, Some Tools → Do the effects we see in a sample probable if applying the hypothesis?

Iterate → toss a coin 250x (140:110)

↳ Count the number of times where H or T occurs as many or more times than the data's statistic.

↳ Use that count and the number of experiments to determine the probability of the statistic happening given the model-based on the fixture not having a real effect.

↳ this is p-value

- p-value ≤ 5 % is by convention stat. sig.
- in practice p-value threshold depends on the test statistic and the model of the null hypothesis

↳ example less 1% is sig.
1%-10% is borderline.
10%+ is insignificant.

p-val = CCDF(x-s).

Null Hypothesis: Pregnancy Length

One tailed versus two-tailed — considering one or both sides of the distribution.

Chi-Squared Test — better for testing proportional differences — rather than using the total deviation.

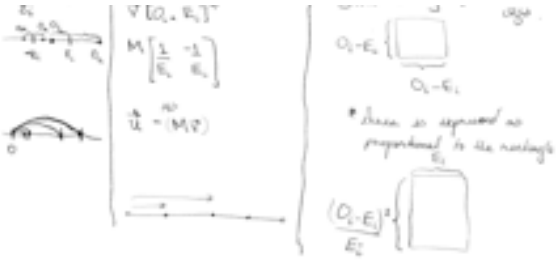
① Testing Hypothesis

- p-value depends on the choice of test statistic, as the test statistic influences the perceived probability of certain experiments.
- The model of the null hypothesis also influences the perceived probability of experiments by making them more or less likely to happen.
- sequencing does weight to largely differing proportions

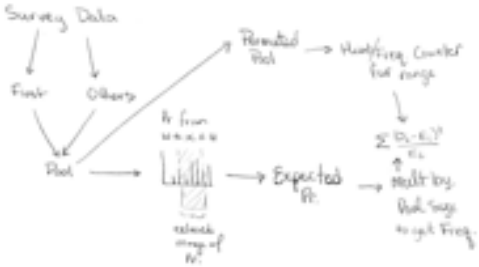
$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Mathematical Intuition

Two-tailed Non-Derivation from | Transformation of $R^2 \to R^1$ | Division by E_i standard



Diagram



ACTUAL
- Test Statistic & Survey Data

EXPERIMENTS
- Test. Stats & Null-Hypothesis Model

} Apply Data

P-Value
- Fraction of Experiments ≥ Actual / Num. of Experiments

Errors Surrounding the P-Value.

False-Positive Rate - probability of wrongly considering something significant.

False-Negative Rate - probability that the hypothesis test will fail when the effect is real!

False Positives occur with an approximate of 1/20 = 5%.

Compute inverse of CDF Right.
- Compute p-val for number of in district. p-val = 1 - CDF(x) = CCDF(x)

- If CCDF(x) ≤ 1/20 → stat. is significant

Since only 5th Percentile of the distribution would qualify as statistical anomalies, that means that if there was other data which would pass the hypothesis test be = 1/20 chance of the data being distributed in way that would then fall in the 5th (1/20) Percentile.

False Negatives

- Depends on the amount of data which represents the effect. - the effect size.
  - ↳ Option #1 - compute false negative rate from

- Take directed [illegible] regions material / moves in sequence, hypothetical [illegible] way (using observed data).
- Go through specifying the REAL Distribution [illegible]
- Get your hands dirty [illegible]
- Read papers until you feel like you're going to [illegible]

Repeat a number of times

- Find a real query to validate your understanding [illegible]

REAL Distribution → Sample → Hypothesis Test → Count false negatives – given the knowledge that the hypothesis-test should test statistical significance

If you know the relationship in your data set as [illegible] [illegible] dataset.

This yields the negative test rate.

The correct positive rate is called the power of the test; or the sensitivity

The complement of the false negative rate — is the correct positive rate

	negative	positive
false	x	y
correct	z	w

A test power of 80+ is considered acceptable — for detecting difference. Below this threshold the test is underpowered.

A negative hypothesis test doesn't imply the absence of a relation, merely the absence of (in the sample) data to prove that relationship.

(*) Since we know the data correctly has a positive correlation; the only other way to classify the result is incorrectly — that is falsely as a negative. The data couldn't be classified correctly as negative nor wrongly/falsely as positive.

Analogy
Like in life when you know/[illegible] the right things, and in your interaction with the world [illegible] produces [illegible] of [illegible] many of those [illegible]

Accuracy of Inference – Replicating Results.

Statistical tools – like hypothesis tests, are inaccurate, and they can not be utilized without the introduction of inaccuracy or loss of precision.

Therefore exploration and then analysis/iterative analysis of the same dataset has inherent bias; notably your findings may be used to construct your [illegible] and your findings are [illegible]/only found within your sample.

ALTERNATIVES · also Dunnett Correction & Tukey's Procedure.

- Split the exploration & testing data partition.
- adjust the p-value threshold to compensate (control the [illegible] – [illegible])
  (i.e necessarily reflect the inaccuracy that comes from the strength of your reliance)
  via Holm-Bonferroni Method, Bonferroni or Šidák.
- result replication
  - → first paper is considered exploratory
  - → the second is confirmatory
- Control False Discovery Rate (FDR) allows type I – rejection of true hypotheses but control proportion of [illegible]

Holm-Bonferroni Method (Šidák is more accurate)

FWER ≤ α $\quad \alpha' = 1-(1-\alpha)^{1/m}$

- Let $H_1, \ldots, H_m$ be a family of null hypotheses
  & $P_1, \ldots, P_m$ be corresponding p-vals
- Order p-vals (lowest-to-highest) $P_{(1)}, \ldots, P_{(m)}$ and label hypoth.
- Given significance level α, let k be the min. index such $P_{(k)} > \frac{\alpha}{m+1-k}$
- Reject hypotheses $H_{(1)} \ldots H_{(k-1)}$ key[$k_{(k)}, m$]
- If $k=1$ keep all
- If no k, reject all.

- Most researchers accept an alpha of 0.05
- Type 1 Error: unnecessarily reject null hypothesis, when there is in the population — [false alarm]
- Type 2 Error: unnecessarily accept null hypothesis when it is in the population. [miss]
- Correction should be applied to p-values
  when two or more statistical analyses have been performed on the same sample data.
  This is due to the increase in family-wise type 1 error rate.
- Family-wise error rate (FWER/ie correlation):
  - the probability of making at least (1) type 1 error
  for a family of tests — statistically a series of tests on data.

  $FWER \leq 1 - (1-\alpha_{FW})^n$

  α_{FW} - alpha level for