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 <u>Null Hypothesis</u>: 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. |of Sample Data (that is data sampled from a Population which you are trying to make inference)

Coin Toss Example

Toss a coin & get H = 140, T = 110.

Create Model 50% H, 50% Tails

→ Do the effects we see in a sample probably appear in a larger population?

1000x → toss a coin 250x (i.e. 140+110)

↳ Count the number of tosses 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 feature not having a real effect

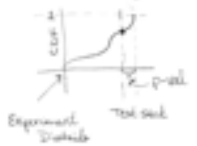
↳ this is p-value

- 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

↳ simple less 1% is sig
1%–10% is borderline.
10%+ is insignificant.

iterations

Simulated data




p-val = CCDF(x - s).

---

Null Hypothesis: Pregnancy Length

Survey Data

First Born $\xrightarrow{\mu}$ μ1st

Other Born $\xrightarrow{\mu}$ μO

Shuffle/ Permute Preg. Len Amongst all data points

First $\xrightarrow{\mu}$ μ1st

Other $\xrightarrow{\mu}$ μO

Repeat n times

① $\boxed{\text{Testing Deviations}}$

A common effect to test is the difference in means between two groups

① Assume the distrib is the same

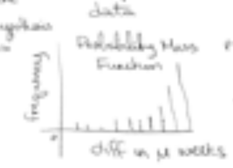
② Create null hypothesis model by permuting the detected values

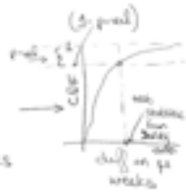
③ Measure probability of the statistic happening based on the data and the null hypothesis model given factor have no effect.

From distribution

Nth experiment data

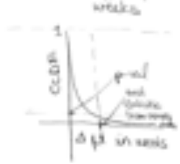
Probability Mass Function




② $\boxed{\text{Testing Correlations}}$

- can also do correlation
  - → Pearson
  - → Spearman's




---

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

---

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

③ $\boxed{\text{Testing Proportions}}$

- 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.
- squaring gives weight to largely differing proportions

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

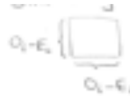

<u>Mathematical Intuition</u>

Two-sided Abs Deviations from | Transformation of ($\mathbb{R}^2 \to \mathbb{R}^2$) Division by $E_i$ Standard

$\nabla[O_i - E_i]^2$

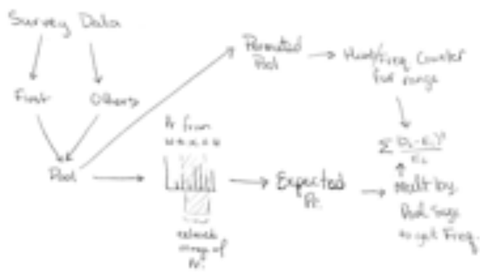
$M\begin{bmatrix} \frac{1}{E_i} & \frac{-1}{E_i} \end{bmatrix}$

$\vec{u} = (M\vec{v})$




* these so expressed as proportional to the rectangle


Diagram




ACTUAL
- Test Statistic of Survey Data

EXPERIMENTS
- Test. Stats of Null-Hypothesis Model

Apply Stats

P-Value
- Fraction of
$\frac{\text{Experiments} \geq \text{Actual}}{\text{Num. of Experiments}}$

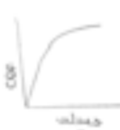
---

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 upperbound of $1/20$ – 5%.




Compute distrib. of Null Hypoth.
- Compute p-val for each x in distrib. 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 valid data which would pass the hypothesis test, it would be a 1/20 chance of the data being distributed in a 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 [illegible] hypothetical [illegible] (using observed data).
  [illegible] material / [illegible] in sequence.
- Go through specifying the model → [illegible]
- Get your hands dirty in [illegible]
- Read papers until you feel like you're going to [illegible]

REAL Distribution

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

Repeat a number of times

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

If you know the relationship in your data set is [illegible], [illegible] correlation dataset.

This yields the negative test rate.

The correct positive rate is called the $\boxed{\text{power}}$ 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/trust [illegible] the [illegible] things, and in your interaction with the world [illegible] produces [illegible] of [illegible] [illegible] 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 maybe used to construct your statistics 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 FWER – [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] to true '+'.

Holm-Bonferroni Method (Šidák is more accurate)

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

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

---

- Most researchers accept an alpha of 0.05.
- Type 1 Error: erroneously reject null hypothesis, when true for the population — $\boxed{\text{the false alarm}}$
- Type 2 Error: erroneously accept null hypothesis when false for the population.
- 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 familywise type 1 error rate.
- Family-wise-error-rate (FWER/α inflation):
  - the probability of making at least (1) type 1 error for a family of tests — essentially a series of tests on data.

$$FWE \leq 1 - (1 - \alpha_{IT})^c$$

$\alpha_{IT}$ — alpha level for