HYPOTHESIS TESTING by Ulysse Zampogna

Framework:

1. **Context Framing**
   - 1.1. Define Business/Product Case
   - 1.2. Define Default Decision
2. **Hypothesis Statements**
   - 2.1. H0 & H1
3. **Success Metrics**
   - 3.1. Primary
   - 3.2. Guardrails
4. **Experiment Design**
   - 4.1. User Journeys
   - 4.2. Minimum Detectable Effect (MDE)
   - 4.3. Significance level
   - 4.4. Statistical Power
   - 4.5. Sample Size
   - 4.6. Experimentation runtime
5. **Decision Making**
   - 5.1. Validity checks
   - 5.2. Statistical test & P-value
   - 5.3. Confidence Intervals
   - 5.4. Final communication & Post launch Monitoring

---

1. Context Framing is the duty of the **decision maker**. It requires business savvy to formulate a product case (which can be backed up by EDA). Then, the decision maker should commit to a default decision up front = "*What am I going to pick if I have to make the decision right now?*". Finally, we can use Statistical inference for decision-making under **uncertainty**.

"The default is the action you're okay with falling into passively whereas the alternative action is something you need to be actively convinced to do."

2. Hypothesis Statements is the art of formulating your ***null hypothesis*** which reflects all possible outcomes where you would take your default decision. The *alternative hypothesis* is the exact opposite (for more complex scenarios, use compound binary hypothesis). H0 & H1 cover all possibilities.

"If data convinces you that you live in the alternative hypothesis world, switch actions."

Check: a Type 1 error should feel worse than a Type 2 error

"The idea of incorrectly leaving your cozy comfort zone (default action) should be more painful than the idea of incorrectly sticking to it."

3. Success Metrics should reflect the main company KPI. The test success depends on 1 primary metric encompassing:

Measurability (Can it be tracked?),
Attributability (Can it be assigned to C/T?),
Sensitiveness (Low variance reach significance more rapidly),
Time (How long shall we wait to report on the success metric?).

Guardrail metric focus on potential negative effects to account for.

4. Experiment Design starts with a thorough documentation of the **user journey** for each treatment - including flow map and screenshots of key differences.

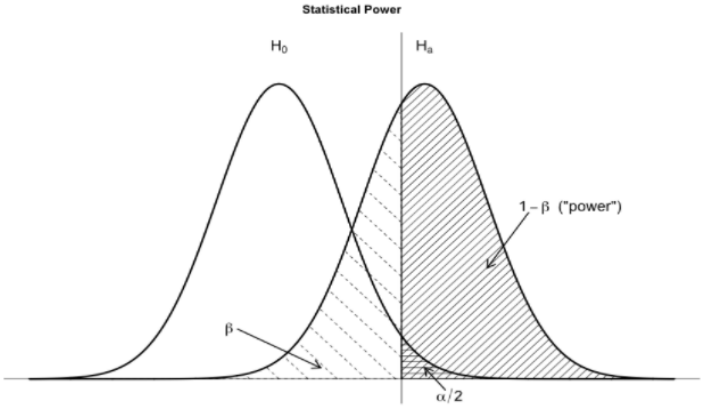
By default, **MDE** can be set at 1%. Otherwise, use an educated guess such as an ROI growth estimate between H0 and H1.

By default, **Significance level** (alpha) is set to 5% or 1%. It's equal to a Type I error (FP) or *P(rejecting H0 | H0 is true)*.

By default, Statistical Power (1-Beta) ranges from 80% to 95%. It's equal to 1-Type II error (FN) or 1-*P(not rejecting H0 | H0 is false)*. In order words,

"With 80% statistical power, you will correctly reject H0 when the effect truly exists 80% of the time. The other 20%, you'll make a Type II error."

How to increase power without increasing alpha? Increase sample size, less variance (depends on metric selection), larger MDE (out of control).



Finally, the **sample size** can be computed thanks to a power analysis (depends on statistical test to be used) utilizing MDE, significance level and statistical power.

Thanks to the sample size, we can derive the **experimentation runtime** depending on ramp up % and daily volume of interest.

5. Decision making : First and foremost, thorough validity checks must be performed across groups such as verifying the randomization, consistent stratification (sampling bias) or checking for potential product interferences.

Statistical tests assume a null hypothesis of no difference between groups. They determine whether the observed data fall outside of the range of values predicted by the null hypothesis.

Choose the type of statistical test depends on:

Verifying important assumptions:
- Independence of observations/variables
- Homogeneity of variance
- Normality of data

Types of variables:
- Continuous
- Discrete (Ordinal, Nominal, Binary)

Parametric tests have stricter requirements and must follow the assumptions cited above. They include **T-test, Z-test or ANOVA**. **Non-Parametric** tests are used when one or more assumptions are violated. They include **Chi Square** among others.

The main output of a statistical test is the p-value - which is the probability of observing the test results observed*, under the assumptions that the null hypothesis is true. P-value is usually set to 5% or 1% and is equal to Type I error.

After clarifying if a statistic of interest is statistically significant, we can report it i.e. mean, median, or another one. We can also report **Confidence intervals** - which is a range of values with a probability that the statistic of interest lies within it. For example, 95% probability the mean lies within 4.2 and 7.3.

6. Final communication : Hypothesis testing is the science of changing your mind. If the evidence collected clearly rejects the null hypothesis and you observed an uplift then action! Implement the alternative hypothesis.

However, we should always start testing by accepting we might learn nothing:

"You should get into the habit of learning nothing more often, because if you insist on learning something beyond the data every time you test hypotheses, you will learn something stupid."

Finally, monitoring key KPIs after launch is crucial.