Section 3F. Hypothesis Testing

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Hypothesis Testing

Let us consider a dataset $\mathcal{D}$ drawn from a linear model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. How can we use the dataset $\mathcal{D}$ to decide whether the input variable X_1 influences the output Y?

- Whenever X_1 does not influence Y, we have that $\beta_1 = 0$ and $Y = \beta_0 + \varepsilon$.
- One might be tempted to answer this question by computing the estimate $\widehat{\beta}_1$ and check if its value is zero.
- However, when $\beta_1 = 0$, we know that $\widehat{\beta}_1 \sim \mathcal{N}(0, \mathrm{SD}(\widehat{\beta}_1)^2)$ and the probability that $\widehat{\beta}_1 = 0$ is zero (why?)

To answer the above question, we state the following two hypotheses:

- A *null hypothesis*, denoted by H_0, stating that there is *no* relationship between X_1 and Y (i.e., $\beta_1 = 0$).
- An *alternative hypothesis*, denoted by H_a, stating that there is *some* relationship between X_1 and Y (i.e., $\beta_1 \neq 0$).

Hypothesis Testing (cont.)

To analyze these hypotheses, we analyze statistical properties of $\widehat{\beta}_1$:

- If H_0 was true ($\beta_1 = 0$), the r.v. $\widehat{\beta}_1$ would follow (approximately) a normal distribution $\mathcal{N}\left(0, \text{SD}\left(\widehat{\beta}_1\right)^2\right)$.
- Hence, with 95% chance, $\widehat{\beta}_1$ is in the CI $\left[-2\text{SD}\left(\widehat{\beta}_1\right), 2\text{SD}\left(\widehat{\beta}_1\right)\right]$.
- In consequence: *if the variable $\widehat{\beta}_1$ is not in the CI, we have strong statistical evidence to reject the null hypothesis!*

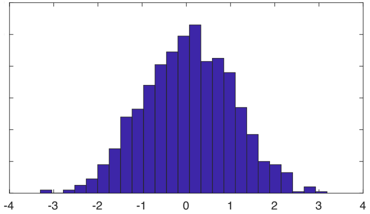
Hypothesis Testing (cont.)

Alternative (but equivalent) tests:

- Assuming that H_0 is true, the random variable $t_1 = \widehat{\beta}_1/\text{SD}\left(\widehat{\beta}_1\right)$ would follow (approximately) a normal distribution $\mathcal{N}(0,1)$.
- Hence, we would have 95% chance that the variable t_1 is in the CI $[-2, 2]$, or $|t_1| \leq 2$.
- In consequence: *if $|t_1| > 2$, then we have strong statistical evidence to claim that H_0 is unlikely to be true!*

Hypothesis Testing: Numerical Illustration

- **Numerical illustration**:
  - Consider a linear model $Y = 0.5 + 0 \cdot X_1 + \varepsilon$, with $f_X \sim \mathcal{N}(0,1)$ and $\sigma^2 = 1$.
  - Generate $1,000$ different datasets $\left\{\mathcal{D}_{\mathrm{Tr}}^k\right\}_{k=1}^{1000}$ using this model and compute $t_1^k = \widehat{\beta}_1^k / \mathrm{SD}(\widehat{\beta}_1^k)$ for each dataset.
  - The histogram below shows the distribution of our estimates $\left\{t_1^k\right\}_{k=1}^{1000}$, which is (approximately) a normal $\mathcal{N}(0,1)$.
  - If the corresponding value of t_1 is *not* in the range $[-2,2]$, we have statistical evidence to reject the null hypothesis $\beta_1 = 0$.



Hypothesis Testing: A Legal Parable

A legal parable: Person A is on trial for murdering person B. Before the trial, the prosecutor collects evidence to prove whether person A has in fact murdered person B. In the legal system of many countries, a person is considered "innocent until proven guilty". Consequently, our null hypothesis H_0 is that A is innocent; the alternative H_a is that A is guilty. A few comments:

- If the prosecutor is able to collect strong evidence to convince the jury that A is guilty (i.e., the null hypothesis is rejected), there is always a chance that A is innocent but convicted unfairly. Similarly, whenever you reject the null hypothesis in a statistical test, there is always a chance that H_0 is true, since our claim is based on statistical evidence (not on direct observation of the ground truth).
- If the prosecutor is not able to convince the jury to declare A "not guilty" (i.e., the null hypothesis is not rejected), that does not imply that A is "innocent". Similarly, you can never prove the null hypothesis to be true using statistical evidence. In other words, "failing to reject" the null hypothesis should not imply "accepting" it.
- In practice, when a person is declared "not guilty", the legal system acts as if the person is innocent. Similarly, in statistics, whenever we fail to reject the null hypothesis, it is common to act as if the null hypothesis is true.

Assessing Model Accuracy

- We define the *Residual Sum-of-Squares (RSS)* as

$$\text{RSS} = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2\,, \text{ where } \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_1$$

  which measures the error between the real outputs y_i and the estimates based on a linear model with intercept $\widehat{\beta}_0$ and slope $\widehat{\beta}_1$

- We also have the *Total Sum-of-Squares (TSS)*

$$TSS = \sum_{i=1}^{N} (y_i - \overline{y})^2\,, \; where\ \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

  Notice that the TSS corresponds to the squared error of a model $Y = \beta_0 + \varepsilon$, where X_1 is assumed to not influence Y

Assessing Model Accuracy (cont.)

- The *R-squared (R^2)* is defined as

  $$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

  Notice that R^2 measures the relative reduction in the squared error before and after we include X_1 as an input variable

- We define the *Residual Standard Error (RSE)* as

  $$\text{RSE} = \sqrt{\frac{\text{RSS}}{N - p - 1}}$$

  the RSE is used as an estimation of the noise standard deviation $\sigma = \sqrt{\text{Var}\left(\varepsilon\right)}$, whenever not available

Penn Engineering