AutoML: Evaluation

Background: Statistical Hypothesis Tests

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Background: statistical hypothesis tests

- When we have a lot of data, we need to summarize it
  - But we already saw that summarization hides a lot of data
  - Ideally, we want to draw high-level conclusions
    (e.g., "A outperforms B on datasets of type X")

Background: statistical hypothesis tests

- When we have a lot of data, we need to summarize it
  - But we already saw that summarization hides a lot of data
  - Ideally, we want to draw high-level conclusions
    (e.g., "A outperforms B on datasets of type X")
- Problem: we only have a finite number of observations
  - Can we attribute observed performance differences to chance?
  - Are we reasonably sure that a claim we make is reproducible?

  $\leadsto$ Statistical tests can help

Statistical hypothesis testing

1. Define initial research hypothesis

Statistical hypothesis testing

1. Define initial research hypothesis
2. Derive null H_0 and alternative H_1 hypothesis
    - Alternative hypothesis should be your research hypothesis

First example: Courtroom Tiral

- A prosecutor tries to prove the guilt of the defendant
- H_0: The defendant is not guilty
  - Accepted for the moment
    ("not guilty as long as their guilt is not proven")
- H_1: The defendant is guilty
  - prosecutor hopes to support that

First example: Courtroom Tiral

- A prosecutor tries to prove the guilt of the defendant
- H_0: The defendant is not guilty
  - Accepted for the moment
    ("not guilty as long as their guilt is not proven")
- H_1: The defendant is guilty
  - prosecutor hopes to support that

First example: Courtroom Tiral

- A prosecutor tries to prove the guilt of the defendant
- H_0: The defendant is not guilty
  - Accepted for the moment
    ("not guilty as long as their guilt is not proven")
- H_1: The defendant is guilty
  - prosecutor hopes to support that

	Truly not guilty	Truly guilty
Found not guilty	Acquittal	Type II Error
Found guilty	Type I Error	Conviction

$\rightsquigarrow$ We want to minimize Type I error!

Statistical hypothesis testing (cont'd)

1. Define initial research hypothesis
2. Derive null H_0 and alternative H_1 hypothesis
   - Alternative hypothesis should be your research hypothesis
3. Consider statistical assumptions
   - E.g., is your data Gaussian distributed?

Statistical hypothesis testing (cont'd)

1. Define initial research hypothesis
2. Derive null H_0 and alternative H_1 hypothesis
   - Alternative hypothesis should be your research hypothesis
3. Consider statistical assumptions
   - E.g., is your data Gaussian distributed?
4. Decide test and test statistic T
   - The correct test depends on your statistical assumptions.
   - Typically: if you use more assumptions, the test is more powerful (i.e., less Type-I error)

Statistical hypothesis testing (cont'd)

1. Define initial research hypothesis
2. Derive null H_0 and alternative H_1 hypothesis
   - Alternative hypothesis should be your research hypothesis
3. Consider statistical assumptions
   - E.g., is your data Gaussian distributed?
4. Decide test and test statistic T
   - The correct test depends on your statistical assumptions.
   - Typically: if you use more assumptions, the test is more powerful (i.e., less Type-I error)
5. Decide significance level α
   (i.e., acceptable Type-I error to reject null hypothesis)

Statistical hypothesis testing (cont'd)

1. Define initial research hypothesis
2. Derive null H_0 and alternative H_1 hypothesis
   - Alternative hypothesis should be your research hypothesis
3. Consider statistical assumptions
   - E.g., is your data Gaussian distributed?
4. Decide test and test statistic T
   - The correct test depends on your statistical assumptions.
   - Typically: if you use more assumptions, the test is more powerful (i.e., less Type-I error)
5. Decide significance level α
   (i.e., acceptable Type-I error to reject null hypothesis)
6. Compute observed t_{obs} of test statistic T
7. Calculate p-value given t_{obs}
   - i.e., probability under the null hypothesis of sampling a test statistic as extreme as observed (probability of Type-I error)

Statistical hypothesis testing (cont'd)

1. Define initial research hypothesis
2. Derive null H_0 and alternative H_1 hypothesis
   - Alternative hypothesis should be your research hypothesis
3. Consider statistical assumptions
   - E.g., is your data Gaussian distributed?
4. Decide test and test statistic T
   - The correct test depends on your statistical assumptions.
   - Typically: if you use more assumptions, the test is more powerful (i.e., less Type-I error)
5. Decide significance level α
   (i.e., acceptable Type-I error to reject null hypothesis)
6. Compute observed t_{obs} of test statistic T
7. Calculate p-value given t_{obs}
   - i.e., probability under the null hypothesis of sampling a test statistic as extreme as observed (probability of Type-I error)
8. If $p < \alpha$, reject null hypothesis in favor of alternative hypothesis
   - If $p > \alpha$, it doesn't tell you anything about the null hypothesis!

Second example for a statistical test

- Claim: “the students in this course are more intelligent than average”

Second example for a statistical test

- Claim: "the students in this course are more intelligent than average"

- Null Hypothesis H_0: $\mu = 100$ (μ is the population mean of this class)
- Alternative Hypothesis H_1: $\mu > 100$ (one-sided hypothesis)

Second example for a statistical test

- Claim: "the students in this course are more intelligent than average"

- Null Hypothesis H_0: $\mu = 100$ (μ is the population mean of this class)
- Alternative Hypothesis H_1: $\mu > 100$ (one-sided hypothesis)

- IQ values are known to be normally distributed with $X \sim \mathcal{N}(100, 15)$
  - $\rightarrow$ statistical assumption

Second example for a statistical test

- Claim: "the students in this course are more intelligent than average"

- Null Hypothesis H_0: $\mu = 100$ (μ is the population mean of this class)
- Alternative Hypothesis H_1: $\mu > 100$ (one-sided hypothesis)

- IQ values are known to be normally distributed with $X \sim \mathcal{N}(100, 15)$
  - $\rightarrow$ statistical assumption

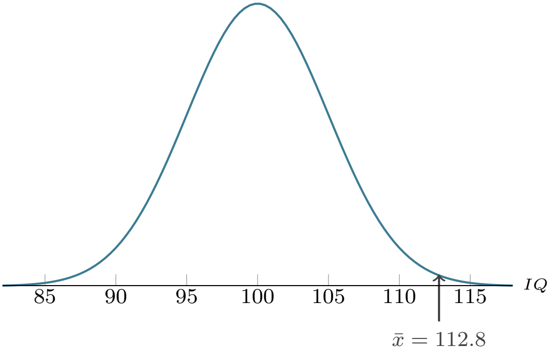
- Let's say we observed IQ values x_i of 9 students in the class:
  - $\{x_1, \ldots, x_9\} = \{116, 128, 125, 119, 89, 99, 105, 116, 118\}$.
  - The sample mean is $\bar{x} = 112.8$
  - Does this data support the claim?

Example continued

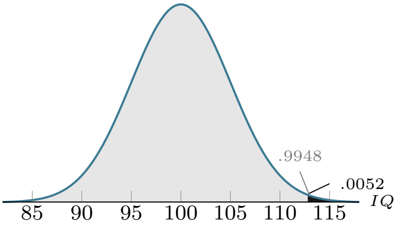
- Distribution of the test statistic
  - Under H_0, we know that each $x_i \sim \mathcal{N}(100, 15)$
  - The test statistic that we measure is the sample mean $\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i$

Example continued

- Distribution of the test statistic
  - Under H_0, we know that each $x_i \sim \mathcal{N}(100, 15)$
  - The test statistic that we measure is the sample mean $\bar{x} = \frac{1}{9} \sum_{i=1}^{9} x_i$
  - Under H_0, the distribution of $\bar{x}$ is $\mathcal{N}(100, 15/\sqrt{9})$
    - Our observation $\bar{x} = 112.8$ is quite extreme under that distribution

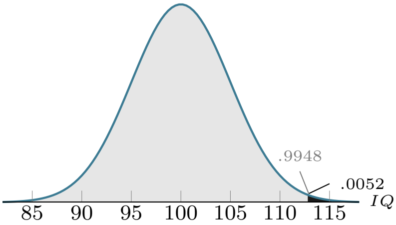


General principle



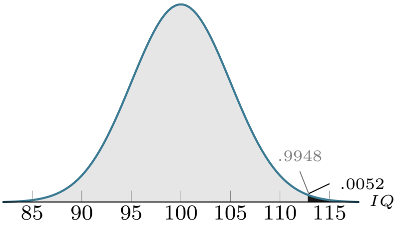
- Compare the test statistic (here: $\bar{x}$) to its sampling distribution under H_0

General principle



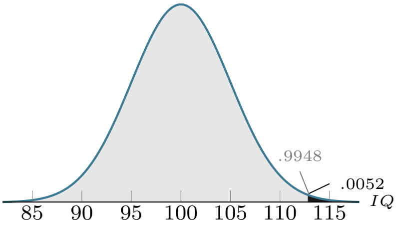
- Compare the test statistic (here: $\bar{x}$) to its sampling distribution under H_0
- P-value: probability p of observing values at least as extreme as $\bar{x}$

General principle



- Compare the test statistic (here: $\bar{x}$) to its sampling distribution under H_0
- P-value: probability p of observing values at least as extreme as $\bar{x}$
- Compare p to pre-defined confidence level α (usually $\alpha = 0.05$); if $p < \alpha$, reject H_0

General principle



- Compare the test statistic (here: $\bar{x}$) to its sampling distribution under H_0
- P-value: probability p of observing values at least as extreme as $\bar{x}$
- Compare p to pre-defined confidence level α (usually $\alpha = 0.05$); if $p < \alpha$, reject H_0
- With $\alpha = 0.01$, would we reject H_0 in this case?

Summary of example

- We just used a so-called Z-test
- H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$
- Assumptions: $X \sim \mathcal{N}(\mu, \sigma^2)$, with known μ and σ^2

Summary of example

- We just used a so-called Z-test
- H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$
- Assumptions: $X \sim \mathcal{N}(\mu, \sigma^2)$, with known μ and σ^2
- Test statistic: sample mean $\bar{x}$; evaluate under $\mathcal{N}(\mu = \mu_0, s = \sigma^2/\sqrt{n})$

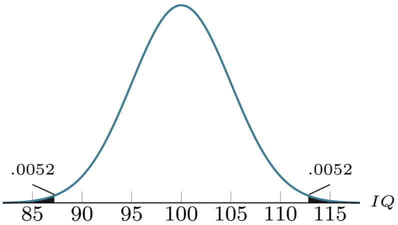
Summary of example

- We just used a so-called Z-test
- H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$
- Assumptions: $X \sim \mathcal{N}(\mu, \sigma^2)$, with known μ and σ^2
- Test statistic: sample mean $\bar{x}$; evaluate under $\mathcal{N}(\mu = \mu_0, s = \sigma^2/\sqrt{n})$
- Equivalent: compute the Z-statistic: $Z = (\bar{x} - \mu_0)/s$ and evaluate cumulative distribution $\Phi(Z)$ of Z under $\mathcal{N}(0, 1)$

Summary of example

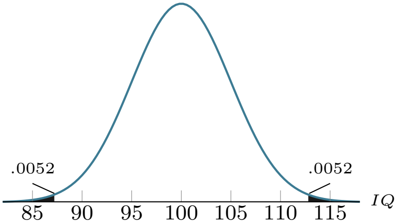
- We just used a so-called Z-test
- H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$
- Assumptions: $X \sim \mathcal{N}(\mu, \sigma^2)$, with known μ and σ^2
- Test statistic: sample mean $\bar{x}$; evaluate under $\mathcal{N}(\mu = \mu_0, s = \sigma^2/\sqrt{n})$
- Equivalent: compute the Z-statistic: $Z = (\bar{x} - \mu_0)/s$ and evaluate cumulative distribution $\Phi(Z)$ of Z under $\mathcal{N}(0,1)$
  - There are standard tables to look up $\Phi(Z)$ for different values of Z
  - Nowadays, there are standard libraries to compute $\Phi(Z)$

Two-sided tests



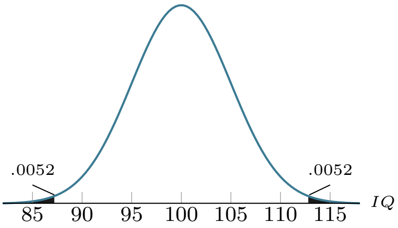
- Similar to one-sided tests, but testing for extreme values in both tails
- Example Z-test: two-sided alternative hypothesis H_1: $\mu \neq \mu_0$

Two-sided tests



- Similar to one-sided tests, but testing for extreme values in both tails
- Example Z-test: two-sided alternative hypothesis H_1: $\mu \neq \mu_0$
- Compute $Z = (\bar{x} - \mu_0)/s$ as before
- Compute p-value as $p = 2\Phi(Z)$, to account for both tails

Two-sided tests



- Similar to one-sided tests, but testing for extreme values in both tails
- Example Z-test: two-sided alternative hypothesis H_1: $\mu \neq \mu_0$
- Compute $Z = (\bar{x} - \mu_0)/s$ as before
- Compute p-value as $p = 2\Phi(Z)$, to account for both tails
- With $\alpha = 0.01$, would we reject H_0 in this case?

General points about statistical hypothesis tests

- What if $p > \alpha$?
  - Failure to reject H_0
  - We cannot conclude that we can accept H_0!

General points about statistical hypothesis tests

- What if $p > \alpha$?
  - Failure to reject H_0
  - We cannot conclude that we can accept H_0!

- Beware (i): most tests make some assumptions
  - E.g., Z-test and popular t-test assume normality
  - Our data is often far from normally-distributed
    - $\rightsquigarrow$ E.g., exponential runtime distributions of optimizers
    - $\rightsquigarrow$ E.g., distribution of fitting a neural network with different random seeds is not well studied

General points about statistical hypothesis tests

- What if $p > \alpha$?
  - Failure to reject H_0
  - We cannot conclude that we can accept H_0!
- Beware (i): most tests make some assumptions
  - E.g., Z-test and popular t-test assume normality
  - Our data is often far from normally-distributed
    - $\leadsto$ E.g., exponential runtime distributions of optimizers
    - $\leadsto$ E.g., distribution of fitting a neural network with different random seeds is not well studied
- Beware (ii): if you use cross-validation, observations are not independent (you cannot apply statistical tests that assume independence)