Confidence Intervals and Hypothesis Testing

Niels Landwehr
Agenda

- Confidence Intervals
- Statistical Tests
Agenda

- Confidence Intervals
- Statistical Tests
Recap: Risk Estimation

- Recap: risk estimation.
- We have learned a model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$.

- Interested in risk of model: the expected loss on novel test instances $(x, y)$ drawn from the data distribution $p(x, y)$.

$$R(\theta) = E[\ell(y, f_\theta(x))] = \int \int \ell(y, f_\theta(x)) p(x, y) dx dy$$

- Because $p(x, y)$ is unknown, risk needs to be estimated from sample $S = (x_1, y_1), \ldots, (x_n, y_n)$ where $(x_i, y_i) \sim p(x, y)$ are independent samples.

- Risk estimate ("empirical risk") $\hat{R}_S(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$

- If context is clear, we denote risk by $R$ and empirical risk by $\hat{R}_S$. 
Recap: Risk Estimation Zero-one loss

- For this lecture, we will assume
  - Learning task is binary classification, $\mathcal{Y} = \{0,1\}$.
  - Loss is zero-one loss,

\[
\ell(y, f_\theta(x)) = \begin{cases} 
0 : y = f_\theta(x) \\
1 : \text{otherwise}
\end{cases}
\]

- This means that $\ell(y_i, f_\theta(x_i))$ for $(x_i, y_i) \sim p(x, y)$ follows a Bernoulli distribution: there is either a mistake or not (coin toss).

- We also assume that model is evaluated on independent test set, such that the error estimate is unbiased.
Idea Confidence Intervals

- Risk estimate is always uncertain – depends on sample $S$.
- Idea confidence interval:
  - Specify interval around risk estimate $\hat{R}_s$
  - Such that the true risk $R$ lies within the interval „most of the time“.
  - Quantifies uncertainty of risk estimate.

Route to confidence interval: analyse the distribution of the random variable $\hat{R}_s$. 

$[ \hat{R}_s ]$

width $\epsilon$ of confidence interval
Central Limit Theorem

- **Central Limit Theorem.** Let \( z_1, \ldots, z_n \) be independent draws from a distribution \( p(z) \) with \( \mathbb{E}[z] = \mu \) and \( \text{Var}[z] = \sigma^2 \). Then it holds that

\[
\sqrt{n} \left( \frac{1}{n} \sum_{j=1}^{n} z_j - \mu \right) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)
\]

average of \( z_1, \ldots, z_n \).

- **Central limit theorem gives approximate distribution of mean:**

\[
\sqrt{n} \left( \frac{1}{n} \sum_{j=1}^{n} z_j - \mu \right) \sim \mathcal{N}(0, \sigma^2)
\]

(approximately, for large \( n \))

\[
\Rightarrow \frac{1}{n} \sum_{j=1}^{n} z_j \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})
\]

(approximately, for large \( n \))
Example Central Limit Theorem

- Example central limit theorem: average of Bernoulli variables.
- Let \( z_1, \ldots, z_n \) be independent draws from a Bernoulli distribution, that is
  \[
  z_i \sim \text{Bern}(z_i \mid \mu) \quad \text{(coin toss with success probability} \ \mu)\]
- Average \( \frac{1}{n} \sum_{j=1}^{n} z_j \) follows (rescaled) Binomial distribution.
- Binomial distribution approaches Normal distribution.
Central Limit Theorem: Error Estimator

- Application of central limit theorem to error estimator.
- Error estimator
  \[ \hat{R}_s = \frac{1}{n} \sum_{j=1}^{n} \ell(y_j, f_\theta(x_j)) \]
  is an average over the Bernoulli-distributed variables \( \ell(y_j, f_\theta(x_j)) \).

- Because the error estimate is unbiased, \( \mathbb{E}[\ell(y_j, f_\theta(x_j))] = R \).
- Variance of Bernoulli random variable is \( \text{Var}[\ell(y_j, f_\theta(x_j))] = R(1 - R) \).

- Central limit theorem says:
  \[ \hat{R}_s \sim \mathcal{N}(R, \frac{R(1-R)}{n}) \]  
  (approximately, large enough \( n \))

- First result for distribution of \( \hat{R}_s \), but depends on \( R \).
Mean and Variance of Error Estimator

- First result: Approximate distribution of error estimator is
  \[ \hat{R}_s \sim \mathcal{N}(R, \frac{R(1-R)}{n}). \]

- Unbiased estimator, therefore the mean is the true risk \( R \).

- The variance of the estimator falls with \( n \): the more instances in the test set \( S \), the less variance.

  - Variance \( \sigma^2_{\hat{R}_s} = \frac{R(1-R)}{n} \).
  
  - Standard deviation ("standard error") \( \sigma_{\hat{R}_s} = \sqrt{\frac{R(1-R)}{n}} \).

Characterizes how much risk estimate fluctuates with \( S \).
Distribution of Error Estimator

- Distribution of error estimator:

\[ \hat{R}_s \sim \mathcal{N}(R, \sigma^2_{\hat{R}_s}). \]

\[ \Rightarrow \frac{\hat{R}_s - R}{\sigma_{\hat{R}_s}} \sim \mathcal{N}(0,1) \]

- Problem: true risk \( R \) has to be known in order to determine variance

\[ \sigma^2_{\hat{R}_s} = \frac{R(1-R)}{n}. \]

- Idea: replace true variance \( \sigma^2_{\hat{R}_s} \) by variance estimate

\[ s^2_{\hat{R}_s} = \frac{\hat{R}_s (1-\hat{R}_s)}{n}. \]
Variance Estimate and t-Distribution

- If true variance is replaced by variance estimate, the normal distribution becomes a Student’s t-distribution:

\[ \frac{\hat{R}_S - R}{s_{\hat{R}_S}} \sim t(n) \]

\( n \) degrees of freedom

- However, for large \( n \) the t-distribution becomes a normal distribution again, so we can keep working with the normal.

Convergence: \( \lim_{n \to \infty} t(n) = \mathcal{N}(0,1) \)
Bound For True Risk

- So what does the empirical risk $\hat{R}_s$ tell us about the true risk?
- From empirical risk $\hat{R}_s$ compute empirical variance $s_{\hat{R}_s}^2$.
- One-sided upper bound for true risk: probability that true risk is at most $\epsilon$ above estimated risk.

$$p(R \leq \hat{R}_s + \epsilon) = p(R - \hat{R}_s \leq \epsilon)$$

$$= p\left(\frac{R - \hat{R}_s}{s_{\hat{R}_s}} \leq \frac{\epsilon}{s_{\hat{R}_s}}\right)$$

$$\frac{\hat{R}_s - R}{s_{\hat{R}_s}} \sim \mathcal{N}(0,1)$$

$$\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(x \mid 0,1)dx$$

"cumulative distribution function of standard normal distribution"
Bound For True Risk

- Symmetric lower bound: because the distribution of $\hat{R}_S$ is symmetric around $R$ (normal distribution), we can similarly compute probability that true risk is at most $\epsilon$ below estimated risk.

  $$p(R \geq \hat{R}_S - \epsilon) \approx \Phi\left(\frac{\epsilon}{s_{\hat{R}_S}}\right)$$

- Two-sided interval: What is the probability that true risk is at most $\epsilon$ away from estimated risk?

  $$p(|R - \hat{R}_S| \leq \epsilon) = 1 - p(R - \hat{R}_S > \epsilon) - p(\hat{R}_S - R > \epsilon)$$

  $$\approx 1 - 2\left(1 - \Phi\left(\frac{\epsilon}{s_{\hat{R}_S}}\right)\right)$$
One-sided and Two-sided Intervals

- So far, we have computed probability that a bound holds for a particular interval size $\varepsilon$.
- Idea: choose $\varepsilon$ in such a way that bounds hold with a certain prespecified probability $1-\delta$ (e.g. $\delta=0.05$).
- One-sided $1-\delta$-confidence interval: bound $\varepsilon$ such that
  $$p(R \leq \hat{R}_s + \varepsilon) = 1 - \delta$$
- Two-sided $1-\delta$-confidence interval: bound $\varepsilon$ such that
  $$p(|R - \hat{R}_s| \leq \varepsilon) = 1 - \delta$$
- For symmetric distributions (here: normal) it always holds that:
  - $\varepsilon$ for one-sided $1-\delta$-interval = $\varepsilon$ for two-sided $1-2\delta$ interval.
  - $\varepsilon$ for one-sided 95%-interval = $\varepsilon$ for two-sided 90% interval.
  - Thus, it suffices to derive $\varepsilon$ for one-sided interval.
Size of Interval

- Compute one-sided $1-\delta$-confidence interval: Determine $\varepsilon$ such that bound holds with probability $1-\delta$.

$$p(R \leq \hat{R}_S + \varepsilon) = 1 - \delta$$

$$\Phi \left( \frac{\varepsilon}{s_{\hat{R}_S}} \right) = 1 - \delta$$

$$\Leftrightarrow \frac{\varepsilon}{s_{\hat{R}_S}} = \Phi^{-1}(1 - \delta)$$

$$\Leftrightarrow \varepsilon = s_{\hat{R}_S} \Phi^{-1}(1 - \delta)$$

Result from Slide 13

- Two-sided confidence interval is $[\hat{R}_S - \varepsilon, \hat{R}_S + \varepsilon]$ (confidence level $1-2\delta$)
Confidence Interval: Example

- **Example:**
  - We have observed an empirical risk of $\hat{R}_s = 0.08$ on $m = 100$ test instances.

- **Compute** $s_{\hat{R}_s} = \sqrt{\frac{0.08 \cdot 0.92}{100}} \approx 0.027$ empirical standard deviation

- Choosing confidence level $\delta = 0.05$ (one-sided level, two-sided will be $2\delta$)

- **Compute** $\varepsilon = s_{\hat{R}_s} \Phi^{-1}(1-\delta) \approx 0.027 \cdot 1.645 \approx 0.045$.

- The confidence interval $[\hat{R}_s - \varepsilon, \hat{R}_s + \varepsilon]$ contains the true risk in 90% of the cases.
Interpretation of Confidence Intervals

- Care should be used when interpreting confidence intervals: the random variable is the empirical risk $\hat{R}_s$ and the resulting interval, not the true risk $R$.

- Correct:

"The probability of obtaining a confidence interval $\varepsilon$ that contains the true risk from an experiment is 95%"

- Wrong:

"We have obtained a confidence interval $\varepsilon$ from an experiment. The probability that the interval contains the true risk is 95%".
Agenda

- Confidence Intervals
- Statistical Tests
Motivation: we have developed a new learning algorithm (Algorithm 1) and compare it to an older algorithm (Algorithm 2) on 10 data sets.

Algorithm 1 seems better (won on 8 data sets, lost on 2).
- But maybe this is just a random result, based on the particular choice of data sets?

Statistical test: rigorous procedure to decide whether it is likely that Algorithm 1 is indeed giving better accuracy.
Statistical Tests: Framework

- Formulate a *null hypothesis* $H_0$.
  - For example, $H_0$ could be „Algorithm 1 and Algorithm 2 perform equally well“.
  - If the observations are very unlikely under $H_0$, we reject it and conclude the alternative hypothesis $H_1$: one algorithm is better.

- Formulate a *test statistic* $T$ that can be computed from data.
  - For example, the observed number of „wins“.

- We will reject the null hypothesis if the test statistic exceeds a threshold $c$.
  - For example, reject if one algorithm wins more than 90 times out of 100.
Statistical Tests: Framework

- Asymmetry in test: we can only reject the null hypothesis, never conclude that it is true.

\[ H_0 \text{ rejected } \implies \text{conclude } H_1. \]
\[ H_0 \text{ not rejected } \implies \text{cannot conclude anything, no new information.} \]

- Possible outcomes of hypothesis testing:

<table>
<thead>
<tr>
<th></th>
<th>( H_0 \text{ rejected} )</th>
<th>( H_0 \text{ not rejected} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 \text{ true} )</td>
<td>Type I error (wrong conclusion, very bad)</td>
<td>no new information but also no error (ok)</td>
</tr>
<tr>
<td>( H_1 \text{ true} )</td>
<td>correct conclusion (good)</td>
<td>Type II error (not enough power, kind of bad)</td>
</tr>
</tbody>
</table>

- Type I error is worst case (publish a study claiming that new drug cures cancer when in fact it does not).
Statistical Tests: More Formally

- More formally, let $\omega \in \Omega$ denote a true parameter of interest (for example, $\omega$ is the probability that Algorithm 1 wins over Algorithm 2 on a randomly drawn data set).

- Let the null hypothesis be $H_0 : \omega \in \Omega_0$ (for example, $H_0 : \omega = 0.5$).
- The alternative hypothesis is $H_1 : \omega \in \Omega_1 = \Omega \setminus \Omega_0$.

- Let $X \in \mathcal{X}$ be the observations (for example, accuracies of algorithms on the multiple data sets).
- Let $T : \mathcal{X} \rightarrow \mathbb{R}$ be the test statistic.

- We reject the null hypothesis $H_0$ (and conclude that the alternative hypothesis $H_1$ is true) if $T(X) > c$. 
Statistical Tests: Size

- **Size** of a test: (maximal) probability of rejecting the null hypothesis when the null hypothesis is true (bad!).
  \[ \alpha = \sup_{\omega \in \Omega_0} p(T > c \mid \omega). \]

- We don’t want Type I errors, so we have to limit \( \alpha \).
- For example, \( \alpha = 0.05 \): formulate test in such a way that there is at most 5% probability of rejecting null hypothesis wrongly.

- Of course, \( \alpha \) depends on \( c \)
  - If we choose \( c \) very large, we are conservative and \( \alpha \) is low.
  - If we choose \( c \) smaller, we are less conservative.
  - Trading Type I for Type II error.
Sign Test

- Sign test: decide whether the medians of two populations differ.
- Motivation: we evaluate two learning algorithms on 10 datasets.

More formally: Let \((a_1, b_1), \ldots, (a_m, b_m) \in \mathbb{R}^2\) be independently sampled as \((a_i, b_i) \sim p(a, b)\).

Let \(\omega = p(a > b) \in [0,1]\) (”probability that Algorithm 1 wins on randomly drawn data set“).

Let \(H_0: \omega = 0.5\), \(H_1: \omega \in [0,1]\setminus\{0.5\}\).
Sign Test

- Sign test: decide whether the medians of two populations differ.
- Motivation: we evaluate two learning algorithms on 10 datasets.

Let \( (x_1, y_1), \ldots, (x_m, y_m) \) (observed accuracies).

Let \( T = \max \left( |\{i | x_i > y_i\}|, |\{i | x_i < y_i\}| \right) \). “#wins of better algorithm”

We will reject the null hypothesis if \( T > c \), that is, if we see more than \( c \) wins of either algorithm.
Sign Test: Distribution under $H_0$

- How do we choose $c$?
- Limit probability of Type I error, given by $\alpha = p(T > c \mid \omega = 0.5)$.

- Because $(a_i, b_i) \sim p(a, b)$ are sampled independently, the logical variable $(a_i > b_i)$ behaves like a coin toss.
- Thus, the probability of seeing $i$ wins for Algorithm 1 is given by a Binomial distribution.
- How likely is it to observe more than $c$ wins (for either algorithm) if $\omega = 0.5$?

$$p(T > c \mid \omega = 0.5) = 2 \sum_{i=c+1}^{m} \text{Bin}_{0.5,m}(i)$$

Probability of seeing extreme #wins under a fair coin toss model.
Sign Test: Distribution under $H_0$

- So \( \alpha = p(T > c \mid \omega = 0.5) \)
  \[
  = 2 \sum_{i=c+1}^{m} \text{Bin}_{0.5,m}(i)
  = 2(1 - \text{BinCDF}_{0.5,m}(c))
  \]

- So far, computed \( \alpha \) for a given threshold \( c \).
- We can ensure any prespecified \( \alpha \) by solving for \( c \):
  \[
  c = \text{BinCDF}_{0.5,m}^{-1}(1 - \alpha / 2).
  \]
- E.g. for \( \alpha = 0.05 \) we set \( c = \text{BinCDF}_{0.5,m}^{-1}(0.975) \).
Sign Test: p-value

- After observing the value $T$ of the test statistics, we can also compute $\alpha$ for the maximum threshold $c=T-1$ that would still reject the null hypothesis. This is called the *p-value*.

$$p = 2(1 - \text{BinCDF}_{0.5,m}(T-1))$$

- The p-value is the smallest $\alpha$ for which the test would reject $H_0$.
- Typically,
  - $p < 0.001$: very sure that $H_0$ can be rejected.
  - $p < 0.01$: sure that $H_0$ can be rejected.
  - $p < 0.05$ reasonably sure that $H_0$ can be rejected.
  - $p < 0.1$ likely that $H_0$ can be rejected.
Sign Test: Example

- Example sign test:

<table>
<thead>
<tr>
<th>Accuracy Algorithm 1</th>
<th>+</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>+</th>
<th>+</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.85</td>
<td>0.76</td>
<td>0.60</td>
<td>0.70</td>
<td>0.95</td>
<td>0.88</td>
<td>0.73</td>
<td>0.89</td>
<td>0.98</td>
<td>0.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy Algorithm 2</th>
<th>+</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>+</th>
<th>+</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.81</td>
<td>0.73</td>
<td>0.61</td>
<td>0.66</td>
<td>0.91</td>
<td>0.89</td>
<td>0.65</td>
<td>0.82</td>
<td>0.97</td>
<td>0.70</td>
</tr>
</tbody>
</table>

- Compute test statistic: \( T=8 \).
- Compute p-value:

\[
p = 2(1 - \text{BinCDF}_{0.5,10}(7)) = 0.1094
\]

- Test would reject null hypothesis for \( \alpha = 0.2 \), but not for \( \alpha = 0.1 \). This is not considered statistically significant.
## Sign Test: Discussion

- **Summary:** sign test can be applied when we have paired data \((a_1, b_1), ..., (a_m, b_m) \in \mathbb{R}^2\) and want to decide if \(p(a > b) \neq 0.5\).

- **Advantages of sign test:**
  - Few assumptions: the \((a_i, b_i)\) only need to be independent.

- **Disadvantages:**
  - Only uses whether \(a_i > b_i\) or \(a_i < b_i\), not the actual values. This discards some information and can make it harder to reject the null hypothesis.
  - Compares medians rather than means: if algorithm is usually slightly better but in some cases much worse, it would be declared the winner.
Two-Tailed Paired t-Test

- Paired t-test: standard test to determine if means between populations differ (example: do risks of two models differ?).

- Let \((a_1, b_1), \ldots, (a_m, b_m) \in \mathbb{R}^2\) be independently sampled from \(p(a, b)\), that is, \((a_i, b_i) \sim p(a, b)\).

- Let \(\delta_i = a_i - b_i\), let \(\Delta = \frac{1}{m} \sum_{i=1}^{m} \delta_i\), and let \(s^2 = \frac{1}{m} \sum_{i=1}^{m} (\delta_i - \Delta)^2\). 
  \[\text{Variance estimate } \delta_i\]

- Let \(\omega = \mathbb{E}[a] - \mathbb{E}[b]\) denote the difference in population means.

- Null hypothesis \(H_0: \omega = 0\), that is, \(\mathbb{E}[a] = \mathbb{E}[b]\).

- Test statistic \(T = \frac{\sqrt{m} \Delta}{s}\), reject if \(T > c\).
Paired t-Test: Probability of Type I Error

- Paired t-test intuition: if null hypothesis $\mathbb{E}[a] = \mathbb{E}[b]$ holds, would expect small $\Delta$ and therefore $T$. Seeing a large (absolute) $T$ is thus very unlikely under the null hypothesis.

- What is the probability of rejecting the null hypothesis when the null hypothesis is true?

$$\alpha = p(T > c \mid \omega = 0)$$
Paired t-Test: Probability of Type I Error

- Distribution of $T$ if $\omega = 0$:
  - Because $\delta_i$ are independent, Central Limit Theorem says:
    $$\frac{\sqrt{m\Delta}}{\sigma} \sim \mathcal{N}(0,1)$$
    zero mean because $\omega = 0$
  - With estimated variance, becomes $t$-distributed:
    $$\frac{\sqrt{m\Delta}}{s} \sim t(m-1)$$

- Thus, test statistic $T$ follows a $t$-distribution.
- Probability that $T$ exceeds $c$:
  $$\alpha = p\left(\frac{\sqrt{m\Delta}}{s} \bigg| \omega = 0 \right) = 2\int_{c}^{\infty} t(x \big| m-1)dx$$

...
Paired t-Test: p-Value

- Formulate using cumulative distribution function:
  \[ \alpha = 2 \int_{c}^{\infty} t(x \mid m-1)dx = 2(1 - \text{tCDF}_{m-1}(c)) \]

- Can again compute a threshold \( c \) for a prespecified \( \alpha \): if we set \( c = \text{tCDF}_{m-1}^{-1}(1 - \alpha / 2) \), we ensure that the Type I error is at most \( \alpha \) (for example, \( \alpha = 0.05 \)).

- For observed value \( T \) of test statistic, we can again compute the \textit{p-value}: the smallest \( \alpha \) for which \( H_0 \) would be rejected.
  \[ p = 2 \int_{T}^{\infty} t(x \mid m-1)dx = 2(1 - \text{tCDF}_{m-1}(T)) \]
Example Paired t-Test

- Example: Comparing the risks of two predictive models.
- We evaluate models \( f_{old} \) and \( f_{new} \) on test set of size \( m = 20 \).
- Let \( \delta_1, ..., \delta_{20} \) be the difference in loss on the different test examples, that is, \( \delta_i = \ell(y_i, f_{old}(x_i)) - \ell(y_i, f_{new}(x_i)) \).

- Compute \( \Delta = \frac{1}{20} \sum_{i=1}^{20} \delta_i \) and \( s^2 = \frac{1}{20} \sum_{i=1}^{20} (\delta_i - \Delta_T)^2 \).

- Let’s say \( \Delta = 0.25 \) and \( s^2 = 0.3026 \)

- Compute \( T = \left| \frac{\sqrt{m\Delta}}{s} \right| = \frac{\sqrt{20} \cdot 0.25}{\sqrt{0.3026}} \approx 2.03 \).

- Compute \( p = 2(1 - tCDF_{m-1}(2.03)) \approx 0.056 \).
- We can reject \( H_0 \) for \( \alpha = 0.1 \), but not for \( \alpha = 0.05 \).
- Weakly significant.
Discussion t-Test

- **Summary**: paired t-test can be applied when we have paired data \((a_1, b_1), \ldots, (a_m, b_m) \in \mathbb{R}^2\) and want to decide if \(E[a] \neq E[b]\).

- **Advantages t-test**
  - Compares means rather than medians (often more adequate).
  - Usually more powerful than sign test.

- **Disadvantages t-test**
  - It critically relies on assuming that the test statistics is t-distributed. This holds in the limit according to central limit theorem, but might not be satisfied for small \(m\).
  - The test can give wrong results when this assumption is not satisfied.
Statistical Tests: Summary and Outlook

- Statistical testing can determine whether observed empirical differences likely indicate true differences between populations.
  - Formulate a null hypothesis.
  - Define a test statistic based on the observations.
  - Reject null hypothesis if observed value for test statistic is very unlikely under null hypothesis.

- Statistical testing is a large field, and many more tests exist
  - Unpaired test, would have to be used when models are evaluated on different test sets.
  - Wilcoxon signed rank test, $\chi^2$-test, …
  - One-tailed vs. two-tailed tests.