Hypothesis testing and statistical decision theory

Lirong Xia

Rensselaer

Nov 7, 2013

Schedule

- Hypothesis testing (definitions)
- Statistical decision theory
  - a more general framework for statistical inference
  - try to explain the scene behind tests

An example

- The average GRE quantitative score of
  - RPI graduate students vs.
  - national average: 558(139)
- Method 1: compute the average score of all RPI graduate students and compare to national average
- End of class

Another example

- Two heuristic algorithms: which one runs faster in general?
- Method 1: compare them on all instances
- Method 2: compare them on a few "randomly" generated instances

Simplified problem: one sample location test

- You have a random variable X
  - you know
    - the shape of X: normal
    - the standard deviation of X: 1
  - you don't know
    - the mean of X
- After observing one sample of X (with value x), what can you say when comparing the mean to 0?
  - what if you see 10?
  - what if you see 2?
  - what if you see 1?

Some quick answers

- Method 1
  - if x>1.645 then say the mean is strictly positive
- Method 2
  - if x<-1.645 then say the mean is strictly negative
- Method 3
  - if x<-1.96 or x>1.96 then say the mean is non-zero
- How should we evaluate these methods?

The null and alternative hypothesis

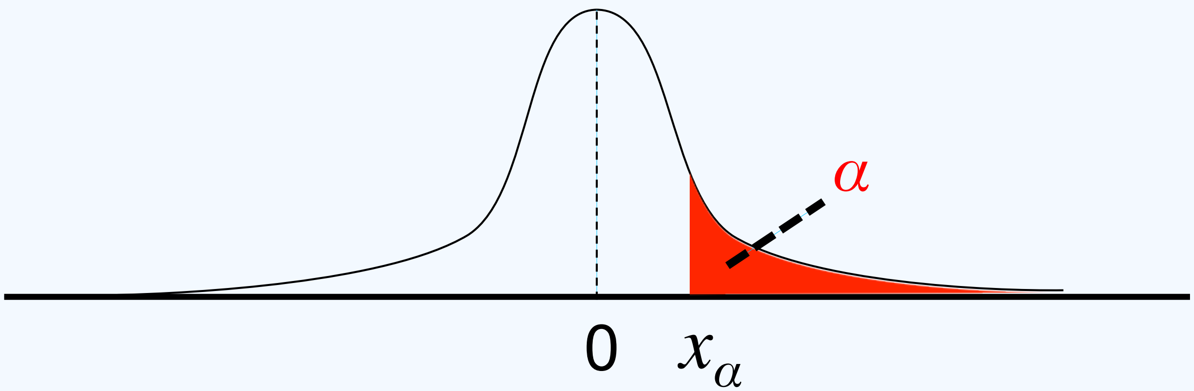
- Given a statistical model
  - parameter space: Θ
  - sample space: S
  - $\Pr(s|\theta)$
- H_1: the alternative hypothesis
  - $H_1 \subseteq \Theta$
  - the set of parameters you think contain the ground truth
- H_0: the null hypothesis
  - $H_0 \subseteq \Theta$
  - $H_0 \cap H_1 = \varnothing$
  - the set of parameters you want to test (and ideally reject)
- Output of the test
  - reject the null: suppose the ground truth is in H_0, it is unlikely that we see what we observe in the data
  - retain the null: we don't have enough evidence to reject the null

One sample location test

- Combination 1 (one-sided, right tail)
  - H_1: mean>0
  - H_0: mean=0 (why not mean<0?)
- Combination 2 (one-sided, left tail)
  - H_1: mean<0
  - H_0: mean=0
- Combination 3 (two-sided)
  - H_1: mean≠0
  - H_0: mean=0
- A hypothesis test is a mapping f : S⟶{reject, retain}

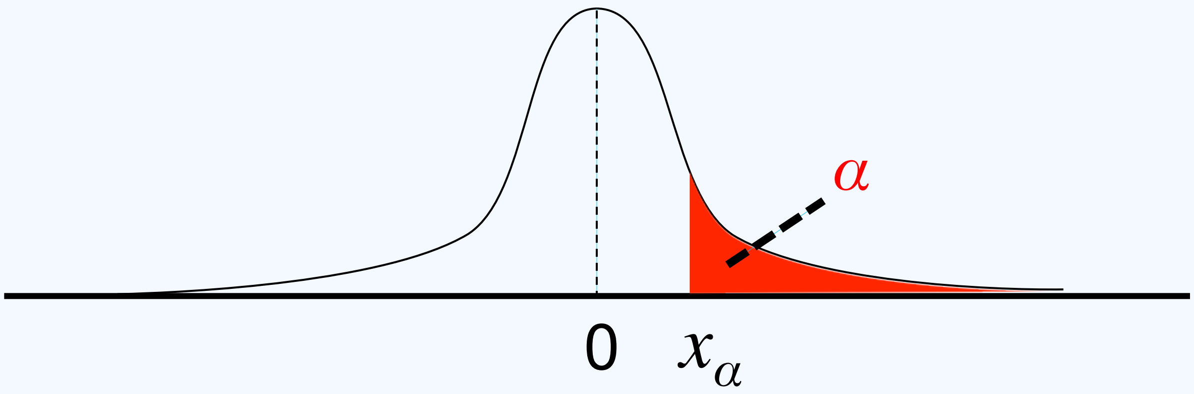
One-sided Z-test

- H_1: mean>0
- H_0: mean=0
- Parameterized by a number $0<\alpha<1$
  - is called the level of significance
- Let x_α be such that $\Pr(X>x_\alpha|H_0)=\alpha$
  - x_α is called the critical value



- Output reject, if
  - $x>x_\alpha$, or $\Pr(X>x|H_0)<\alpha$
    - $\Pr(X>x|H_0)$ is called the p-value
- Output retain, if
  - $x\leq x_\alpha$, or p-value$\geq\alpha$

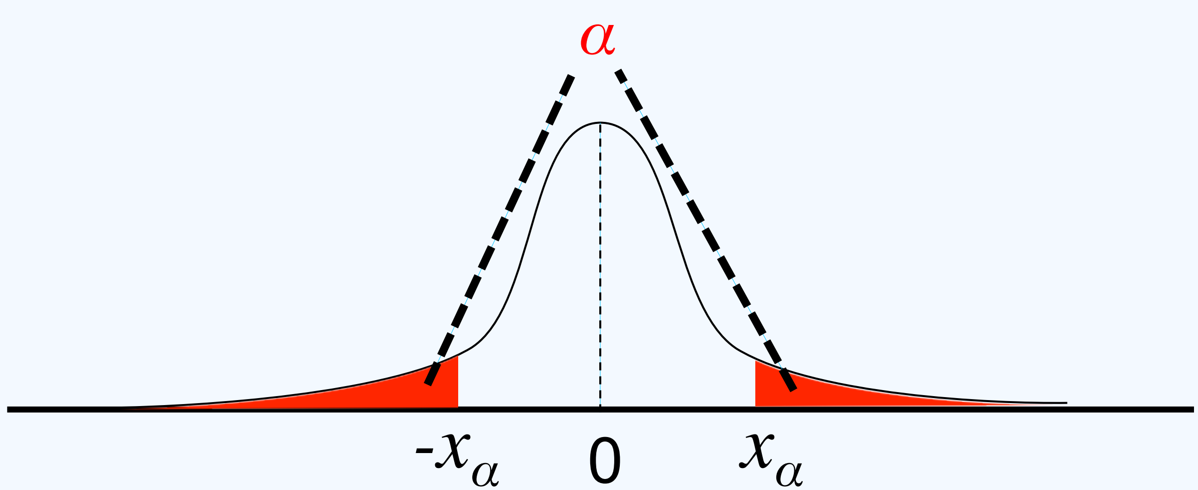
Interpreting level of significance



- Popular values of α:
  - 5%: x_α= 1.645 std (somewhat confident)
  - 1%: x_α= 2.33 std (very confident)
- α is the probability that given mean=0, a randomly generated data will leads to "reject"
  - Type I error

Two-sided Z-test

- H_1: mean≠0
- H_0: mean=0
- Parameterized by a number $0<\alpha<1$
- Let x_α be such that $2\Pr(X>x_\alpha|H_0)=\alpha$



- Output reject, if
  - $x>x_\alpha$, or $x<x_\alpha$
- Output retain, if
  - $-x_\alpha \leq x \leq x_\alpha$

What we have learned so far…

- One/two-sided Z test: hypothesis tests for one sample location test (for different H_1's)
- Outputs either to "reject" or "retain" the null hypothesis
- And defined a lot of seemingly fancy terms on the way
  - null/alternative hypothesis
  - level of significance
  - critical value
  - p-value
  - Type I error

Questions that haunted me when I first learned these

- What the heck are you doing by using different H_1?
  - the description of the tests does not depend on the selection of H_1
  - if we reject H_0 using one-sided test (mean>0), shouldn't we already be able to say mean≠0? Why need two-sided test?
- What the heck are you doing by saying "reject" and "retain"
  - Can't you just predict whether the ground truth is in H_0 or H_1?

Evaluation of hypothesis tests

- What is a "correct" answer given by a test?
  - when the ground truth is in H_0, retain the null (≈saying that the ground truth is in H_0)
  - when the ground truth is in H_1, reject the null (≈saying that the ground truth is in H_1)
  - only consider cases where $\theta \in H_0 \cup H_1$
- Two types of errors
  - Type I: wrongly reject H_0, false alarm
  - Type II: wrongly retain H_0, fail to raise the alarm
  - Which is more serious?

Type I and Type II errors

		Output	
		Retain	Reject
Ground truth in	H_0	size: 1-α	Type I: α
	H_1	Type II: β	power: 1-β

- Type I: the max error rate for all $\theta \in H_0$

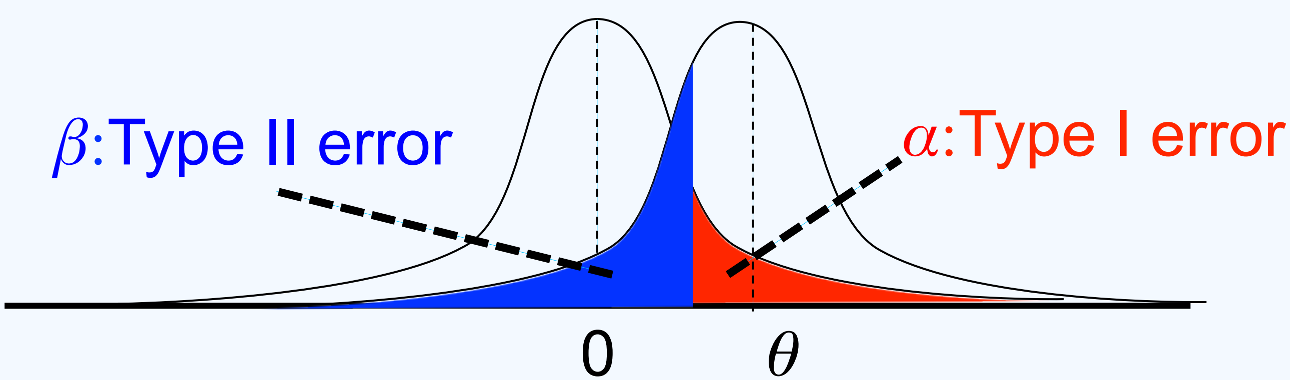
$$\alpha = \sup_{\theta \in H_0} \Pr(\text{false alarm} | \theta)$$

- Type II: the error rate given $\theta \in H_1$
- Is it possible to design a test where $\alpha = \beta = 0$?
  - usually impossible, needs a tradeoff

Illustration

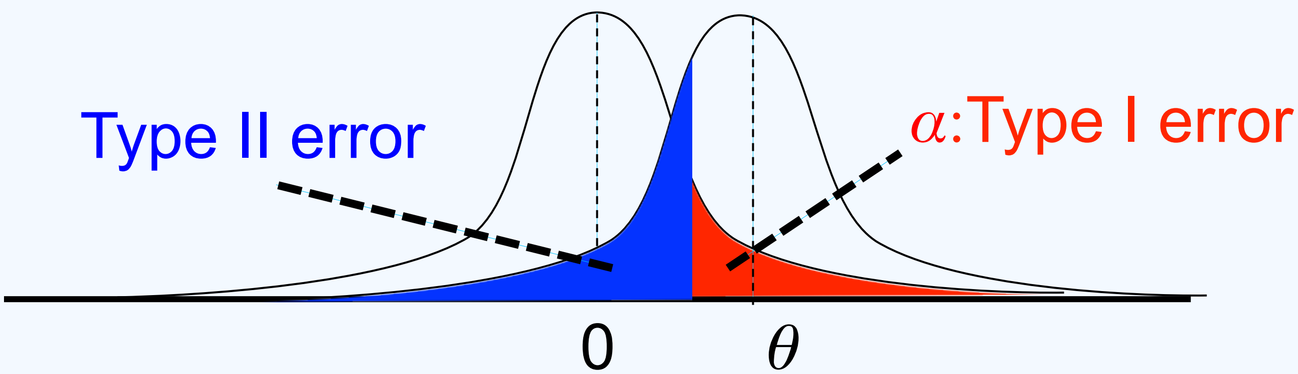
- One-sided Z-test
  - we can freely control Type I error
  - for Type II, fix some $\theta \in H_1$

		Output	
		Retain	Reject
Ground truth in	H_0	size: 1-α	Type I: α
	H_1	Type II: β	power: 1-β

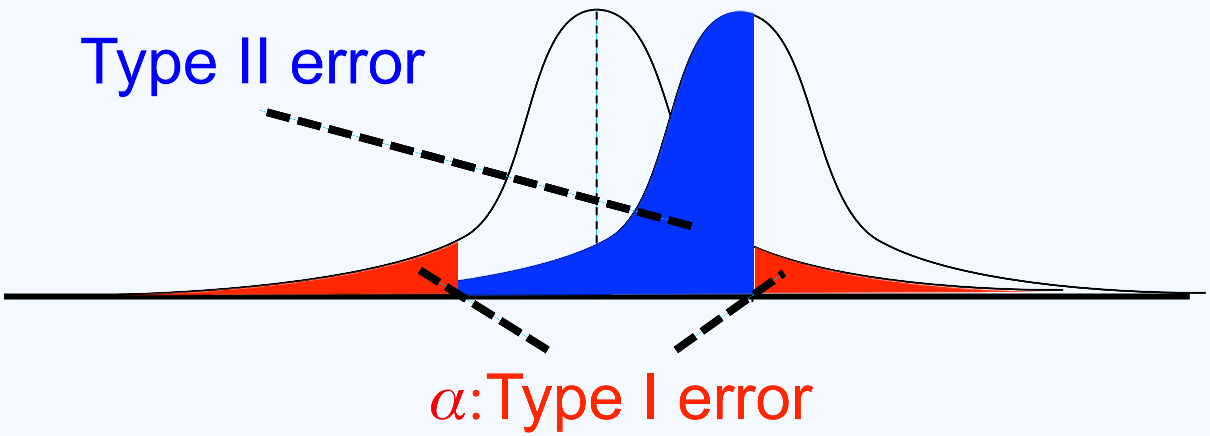


Using two-sided Z-test for one-sided hypothesis

- Errors for one-sided Z-test



- Errors for two-sided Z-test, same α



Using one-sided Z-test for a set-valued null hypothesis

- H_1: mean>0
- H_0: mean≤0 (vs. mean=0)
- $\text{Sup}_{\theta \leq 0}$Pr(false alarm|$\theta$)=Pr(false alarm| $\theta=0$)
  - Type I error is the same
- Type II error is also the same for any $\theta>0$
- Any better tests?

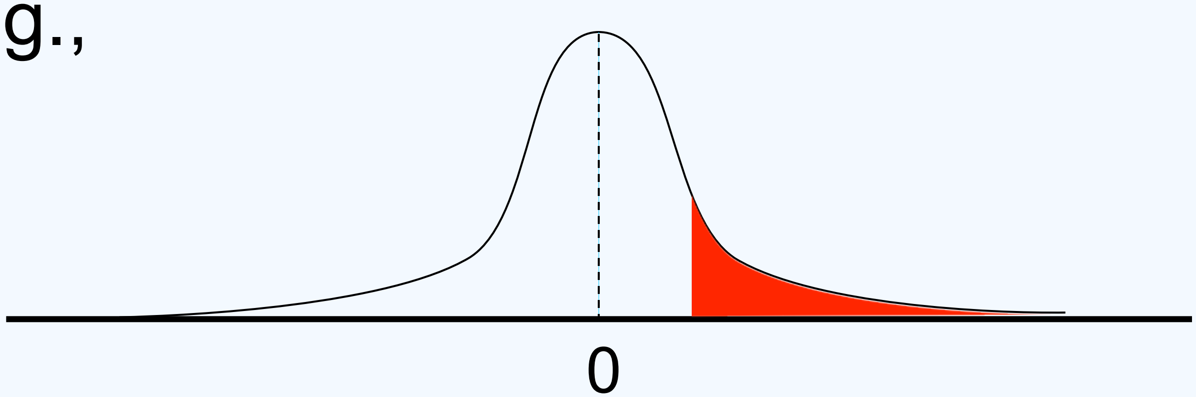
Optimal hypothesis tests

- A hypothesis test f is uniformly most powerful (UMP), if
  - for any other test f' with the same Type I error
  - for any $\theta \in H_1$, we have

    Type II error of f < Type II error of f'
- Corollary of Karlin-Rubin theorem: One-sided Z-test is a UMP for H_0:≤0 and H_1:>0
  - generally no UMP for two-sided tests

Template of other tests

- Tell you the H_0 and H_1 used in the test
  - e.g., H0:mean≤0 and H1:mean>0
- Tell you the test statistic, which is a function from data to a scalar
  - e.g., compute the mean of the data
- For any given α, specify a region of test statistic that will leads to the rejection of H_0
  - e.g.,



How to do test for your problem?

- Step 1: look for a type of test that fits your problem (from e.g. wiki)
- Step 2: choose H_0 and H_1
- Step 3: choose level of significance α
- Step 4: run the test

Statistical decision theory

- Given
  - statistical model: Θ, S, Pr($s|\theta$)
  - decision space: D
  - loss function: L(θ, d)$\in \mathbb{R}$
- We want to make a decision based on observed generated data
  - decision function f : data$\longrightarrow$D

Hypothesis testing as a decision problem

- D={reject, retain}
- L(θ, reject)=
  - 1, if $\theta \in H_1$
  - 0, if $\theta \in H_0$ (type I error)
- L(θ, retain)=
  - 1, if $\theta \in H_0$
  - 0, if $\theta \in H_1$ (type II error)

Bayesian expected loss

- Given data and the decision d
  - $\mathrm{EL_B}(\mathrm{data}, d) = \mathrm{E}_{\theta|\mathrm{data}} \mathrm{L}(\theta, d)$
- Compute a decision that minimized EL for a given the data

Frequentist expected loss

- Given the ground truth θ and the decision function f
  - $EL_F(\theta, f) = E_{data|\theta} L(\theta, f(\theta))$
- Compute a decision function with small EL for all possible ground truth
  - c.f. uniformly most powerful test: for all $\theta \in H_1$, the UMP test always has the lowest expected loss (Type II error)

Recap

- Problem: make a decision based on randomly generated data
- Z-test
  - null/alternative hypothesis
  - level of significance
  - reject/retain
- Statistical decision theory framework
  - Bayesian expected loss
  - Frequentist expected loss