CS145: Probability & Computing

Lecture 19: Frequentist Hypothesis Tests, Bayesian Hypothesis Tests

Brown University Computer Science

Figure credits:
*Bertsekas & Tsitsiklis, **Introduction to Probability**, 2008*
*Pitman, **Probability**, 1999*
Credits:
*Hui Wang, **Hypothesis Testing**, 2017*

From Probability to Statistics

- In *probability theory* we compute the probability that 20 independent flips of a fair (unbiased) coin give the sequence

  HTTHTHTHHTTHTHTHHTTT

- In *statistics* we ask: Given that we observed the sequence

  HTTHTHTHHTTHTHTHHTTT

  what is the likelihood that the coin is fair (unbiased)?

https://hattonsoflondon.co.uk/

CS145: Lecture 20 Outline

- Frequentist Hypothesis Tests
- Bayesian Hypothesis Tests

Bayesian vs Frequentist approach

- **Frequentist:**
  - *Fixed:* The true (but unknown) state of the hypothesis in the world.
  - *Random:* The data, over many hypothetical repetitions of experiment.

Does the data provide enough evidence to reject a null-hypothesis with confidence?

- **Bayesian:**
  - *Fixed:* The single data set we have observed.
  - *Random:* The true value of the hypothesis, given our partial knowledge.

What is the hypothesis which is most likely to be correct?

Hypothesis testing and coinflips

Over Spring 2009 two Berkeley undergraduates, Priscilla Ku and Janet Larwood, undertook a task to perform 40,000 coin tosses.

It was “only” one hour per day for a semester....

Result:

Heads = 20217 times.
Tails = 19783 times.

This outcome could be the result of either fair or biased coin

Question: Is the coin fair?

Hypothesis Testing - Intuition

Question: Is the coin fair?

Define a test BEFORE you run the experiment:
Choose a set of outcomes that is unlikely for a fair coin.
For example, if X is the number of heads in 40,000-coin tosses:

$$Pr(|X - 20,000| \geq 200) \leq 0.05$$

Decision Rule: If $|X - 20,000| \geq 200$ we'll reject the hypothe that the coin is fair.

We now run the test.
We get:
Heads = 20217 times.
Tails = 19783 times.

Using this decision rule, we reject
The hypothesis that the coin is fair

Hypothesis Testing - Intuition

Question: Is the coin fair?

Before you run the experiment:

Define an outcome that is unlikely for a fair coin.

For example, X is the number of heads in 40,000-coin tosses:

$$Pr(|X - 20,000| \geq 260) \leq 0.01$$

Decision rule: if $|X - 20,000| \geq 260$ we'll say that it is "unlikly" that the coin is fair.

We now run the test.

We get:

Heads = 20217 times.

Tails = 19783 times.

With this criteria we cannot reject the hypothesis that the coin is fair

This test requires stronger evidence to decide that the coin is not fair

Hypothesis testing steps

- Formulate your theory “in a testable way”
  - Null Hypothesis
  - Alternative Hypothesis
- Identify your test
  - Test statistics
- Identify how certain you want to be
  - Level of Significance
- Decision criteria
  - Identify a “rejection” region
  - p-value

What is an hypothesis

- A hypothesis is a claim (assumption) about a population parameter (not the observed data):

  - population mean

    Example: The mean monthly cell phone bill of this city is $\mu = \$42$

  - population proportion

    Example: The proportion of adults in this city with cell phones is $p = 68$

The null hypothesis, H_0

Usually refers to the default position

- New theory does not give better explanation
- New medication is not performing better
- …

Hypothesis testing is not symmetric. It gives priority to the null.

The null hypothesis is rejected only if the data shows that it is very unlikely, otherwise the null holds.

The null hypothesis, H_0

States the assumption (numerical) to be tested

Example: the coin is fair $H_0 : p = 0.5$

where p is the probability of head

Is always about a population (or data distribution) parameter, NOT about a sample statistic

$H_0 : p = 0.5$ 

$H_0 : \hat{p} = 0.5$ 

Null hypothesis

We need to decide regarding our coin…

- H_1: The alternative hypothesis - the coin is weighted
- H_0: The null hypothesis - the coin is fair

Researchers do not know which hypothesis is true.
They must make a decision on the basis of evidence presented.

Basic Frequentist Idea

- Hypotheses are fixed: they synthetize a prior belief on the data

- Data is random: the analyst evaluates if the hypothesis is coherent with respect to the random data

The hypothesis testing set-up

1. Set Up *Null Hypothesis (H_0)* and *Alternative Hypothesis (H_a)*:
   For the coin test: H_0: $p = 0.5$, H_a: $p \neq 0.5$

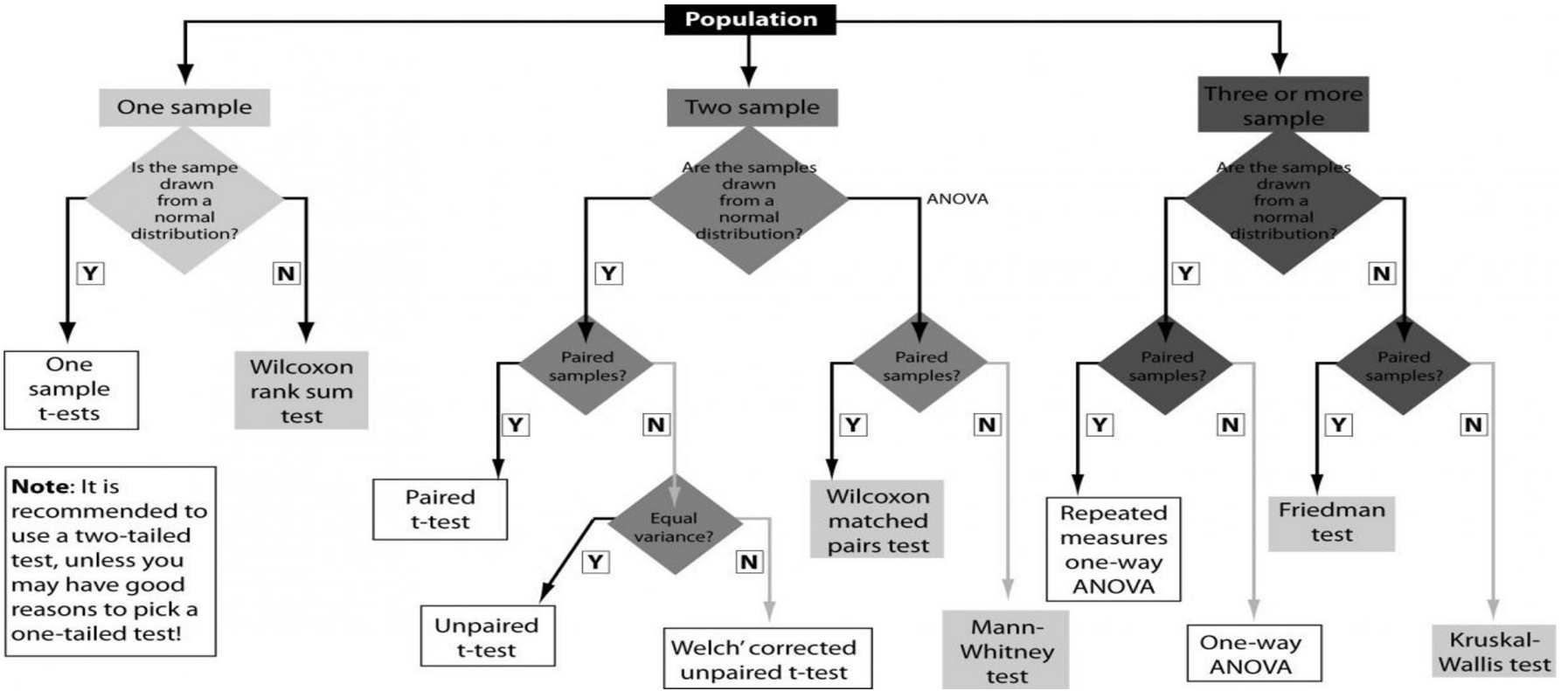
2. Find a Test Statistic: a function of the data.
For the coin test we can use the empirical frequency.

3. Define a rejection criteria – a set of values of the test statistics that is unlikely under the null hypothesis.
For the coin test we can choose $|\hat{p} - 20{,}000| > 200$.

Select your test

- Testing is a bit like finding the right recipe based on these ingredients:
  - Question
  - Data type
  - Sample size
  - Variance known? Variance of several groups equal?
- Good news: Plenty of tables available, e.g.,
  - http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm (with examples in R, SAS, Stata, SPSS)

How to choose your test



Example of a table of tests

Summary Table for Statistical Techniques

Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
Estimating a Mean	One Population Mean μ	Sample mean $\bar{y}$	Numerical	• What is the average weight of adults? • What is the average cholesterol level of adult females?	1-sample t-interval $\bar{y} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	Stat >Basic statistics >1-sample t	• data approximately norm or • have a large sample size ($n \geq 30$)
Test about a Mean	One Population Mean μ	Sample mean $\bar{y}$	Numerical	• Is the average GPA of juniors at Penn State higher than 3.0? • Is the average Winter temperature in State College less than 42° F?	$H_o: \mu = \mu_o$ $H_a: \mu \neq \mu_o$ or $H_a: \mu > \mu_o$ or $H_a: \mu < \mu_o$ The one sample t test: $t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$	Stat >Basic statistics >1-sample t	• data approximately norm or • have a large sample size ($n \geq 30$)
Estimating a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	• What is the proportion of males in the world? • What is the proportion of students that smoke?	1-proportion Z-interval $\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	Stat >Basic statistics >1-sample proportion	• have at least 5 in each category
Test about a Proportion	One Population Proportion π	Sample Proportion $\hat{\pi}$	Categorical (Binary)	• Is the proportion of females different from 0.5? • Is the proportion of students who fail Stat 500 less than 0.1?	$H_o: \pi = \pi_o$ $H_a: \pi \neq \pi_o$ or $H_a: \pi > \pi_o$ or $H_a: \pi < \pi_o$ The one proportion Z-test: $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Stat >Basic statistics >1-sample proportion	• $n\pi_o \geq 5$ and $n(1-\pi_o) \geq$

http://sites.stat.psu.edu/~ajw13/stat500_su_res/notes/lesson14/images/summary_table.pdf

Hypothesis testing with Confidence Level

What Is Significance (Confidence) Level α?

The Confidence Level is the probability that the test data satisfies the rejection rule when H_0 is correct. (The probability of rejection the null when the null is true)

The most common value of α is 5%, though α = 1% is also widely used.

In the coin test we had:

$$Pr(|X - 20,000| \geq 200) \leq 0.05$$

Decision Rule: If $|X - 20,000| \geq 200$ we'll reject the hypothe that the coin is fair.

The confidence level for that decision rule (= rejection set) in 0.05

Hypothesis testing with p-value

What Is p-value?

Instead of fixing the confidence level we can ask what is the minimum confidence level of a test that rejects the null. This is the p-value.

It is the probability of observing test statistics that are as extreme or more extreme than the present empirical data, assuming H_0 is valid.

In the coin case:

$$p - value = Pr\left(\left|\hat{p} - \frac{1}{2}\right| \geq \frac{20217}{40000} - \frac{1}{2}\right)$$

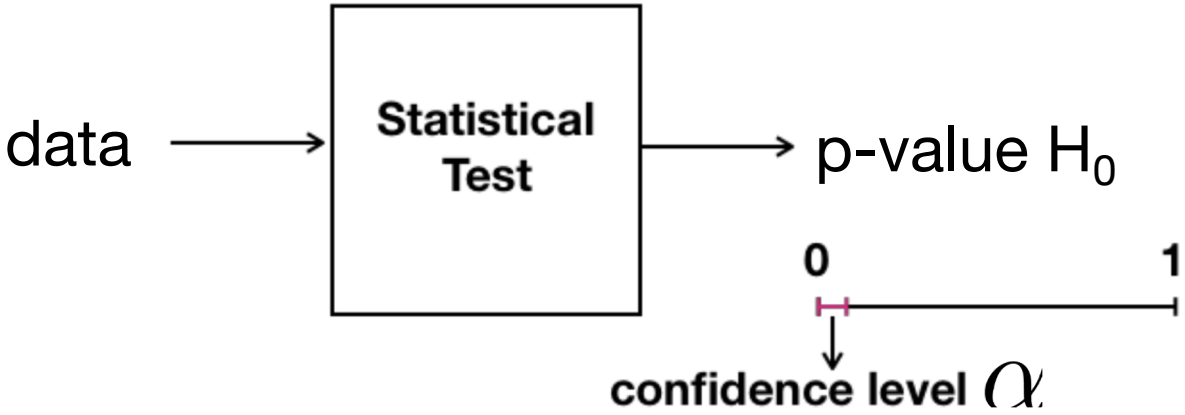
$$= Pr\left(\left|\hat{p} - \frac{1}{2}\right| \geq 0.005425\right) \quad = 0.03$$

The P-value depends on the specific assumptions/test being used!

Null hypothesis is rejected if and only if the P-value is less than the significance level α.

Why is this working?

If H_0 is correct (i.e., should NOT be rejected) then its p-value is uniformly distributed in [0,1]



Hence, P(p-value $H_0 \leq \alpha$ | H_0 is a true null) $= \alpha$

Thus, P(H_0 rejected | H_0 is a true null) $= \alpha$

Comments on hypothesis testing

- Relation Between P-Value and Significance Level α: Null hypothesis is rejected if and only if P-value is less than the significance level α.

- Rejection and Acceptance: Rejection of null hypothesis does not mean null hypothesis is wrong. It means null hypothesis is statistically implausible. Similarly, acceptance of null hypothesis does not mean is correct. It means null hypothesis is not statistically implausible.

- Statistical Significance: Statistical significance is not practical significance — recall the 40000 coin tosses. A small practical discrepancy can be statistically very significant, especially with large data set!

Types of error

Outcome (Probability)

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject **H_0**	**No error** **(1 - α)**	**Type II Error** **(β)**
Reject **H_0**	**Type I Error** **(α)**	**No Error** **(1 - β)**

Types of error and Power of Test

- Type I Error (False Positive): Given a significance level α, what is the chance that null hypothesis will be rejected, even when it is indeed correct?

  Answer: $P(H_0 \text{ rejected } | H_0 \text{ is true}) = \alpha$

- Type II Error (False Negative): Given a significance level α, what is the chance that null hypothesis will be accepted, even when it is indeed wrong?

  Answer: $\beta = 1 - P(H\_0 \text{ is rejected} | H_1 \text{ is true})$

The ideal scenario is that both α and β are small. But they are in conflict! Everything else being equal, one cannot reduce type I error and type II error simultaneously.

- Power of Test: It is defined to be $1 - \beta = P(H\_0 \text{ is rejected} || H_1 \text{ is true})$

Avoiding False Positives

- Usually we are looking for sufficient evidence to reject H_0.
- Type I errors are implicitly more important than type II errors.
- One usually controls type I error below some prefixed small threshold, and then, subject to this control, look for a test which maximizes power or minimizes type II error.

Testing means of normals

- Let $\{X_1, \ldots, X_n\}$ be iid samples from $N(\mu, \sigma^2)$, where σ^2 is known but μ unknown. Want to perform hypothesis testing on μ.

- We consider three scenarios.

  - One-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$
  - One-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu < \mu_0$
  - Two-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$

Testing means of Normals

- Let $\{X_1, \ldots, X_n\}$ be iid samples from $N(\mu, \sigma^2)$, where σ^2 is known but μ unknown. Want to perform hypothesis testing on μ.

- We consider three scenarios.

  - Upper (Right) Tailed Test: $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$
  - Lower (Left) Tailed Test: $H_0 : \mu = \mu_0$, $H_a : \mu < \mu_0$
  - Two-Tailed Test: $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$

Testing means of normals: upper tailed test

- Null hypothesis $H_0 : \mu = \mu_0$, Alternative hypothesis $H_a : \mu > \mu_0$

1. Test statistic : sample mean $\bar{X}$. It is a Random variable!
   We denote as $\bar{x} \sim \bar{X}$ a realization, that is the *observed sample mean*

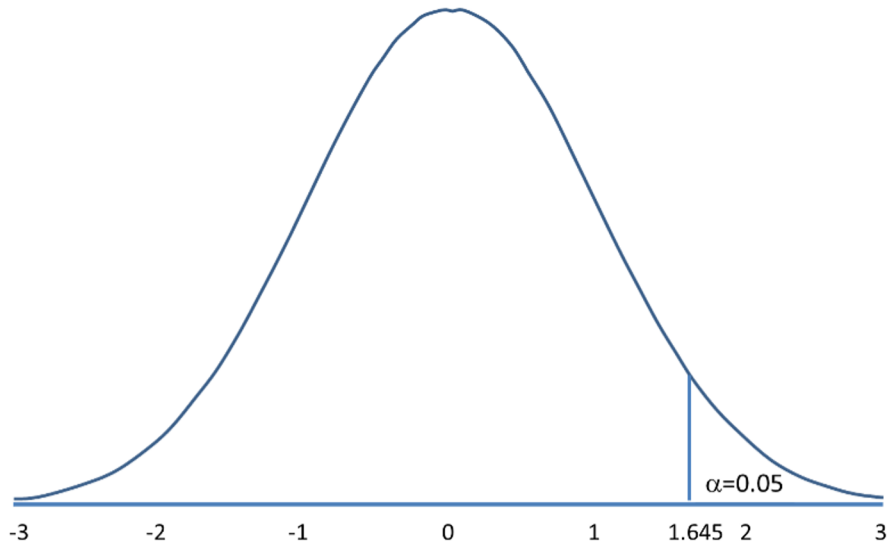
2. p-value computation: under the null-hypothesis $\bar{X} \sim N(\mu_0, \sigma^2/n)$

$$p - value = P(\bar{X} \geq \bar{x}) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(-\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

   where $\Phi()$ denotes the cumulative distribution function of the normal

3. decision: given significance level α, we reject H_0 iff $\alpha \geq p - value$

Upper (right) tailed test



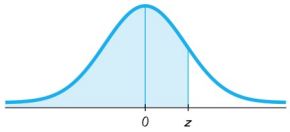
The decision rule is: Reject H_0 if $Z \geq 1.645$.

Upper-Tailed Test	
α	Z
0.10	1.282
0.05	1.645
0.025	1.960
0.010	2.326
0.005	2.576
0.001	3.090
0.0001	3.719

Use the appropriate table!

Use tables for the Standard Normal Distribution (z-tables)

Report the cumulative area from the LEFT

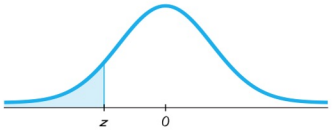


POSITIVE *z* Scores

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

EGATIVE *z* Scores



2 Standard Normal (*z*) Distribution: Cumulative Area from the LEFT

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	.0001									
	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084

Example

NEGATIVE *z* Scores

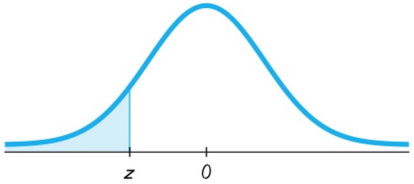


TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084

Suppose we need $\Phi(-2.45)$

Testing means of normals: lower tailed test

- Null hypothesis $H_0 : \mu = \mu_0$, Alternative hypothesis $H_a : \mu < \mu_0$

1. Test statistic : sample mean $\bar{X}$. It is a Random variable!
   We denote as $\bar{x} \sim \bar{X}$ a realization, that is the *observed sample mean*

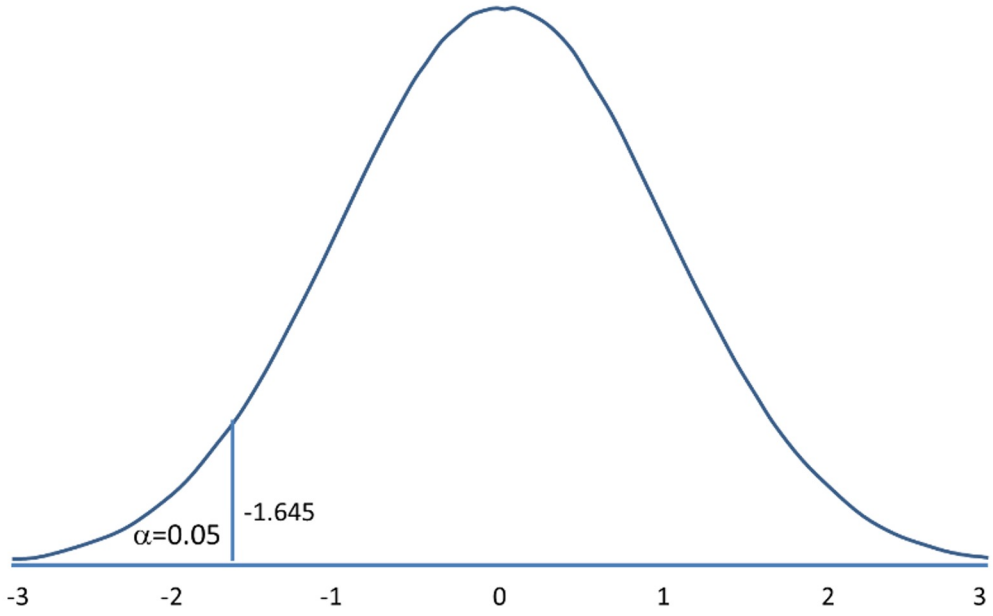
2. p-value computation: under the null-hypothesis $\bar{X} \sim N(\mu_0, \sigma^2/n)$

$$p - value = P(\bar{X} \leq \bar{x}) = \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

   where $\Phi()$ denotes the cumulative distribution function of the normal

3. decision: given significance level α, we reject H_0 iff $\alpha \geq p - value$

Low (left) tailed test



The decision rule is: Reject H_0 if $Z \leq 1.645$.

Lower-Tailed Test	
a	Z
0.10	-1.282
0.05	-1.645
0.025	-1.960
0.010	-2.326
0.005	-2.576
0.001	-3.090
0.0001	-3.719

Testing means of normals: two sided test

- Null hypothesis $H_0 : \mu = \mu_0$, Alternative hypothesis $H_a : \mu \neq \mu_0$

1. Test statistic : sample mean $\bar{X}$. It is a Random variable!
   We denote as $\bar{x} \sim \bar{X}$ a realization, that is the *observed sample mean*

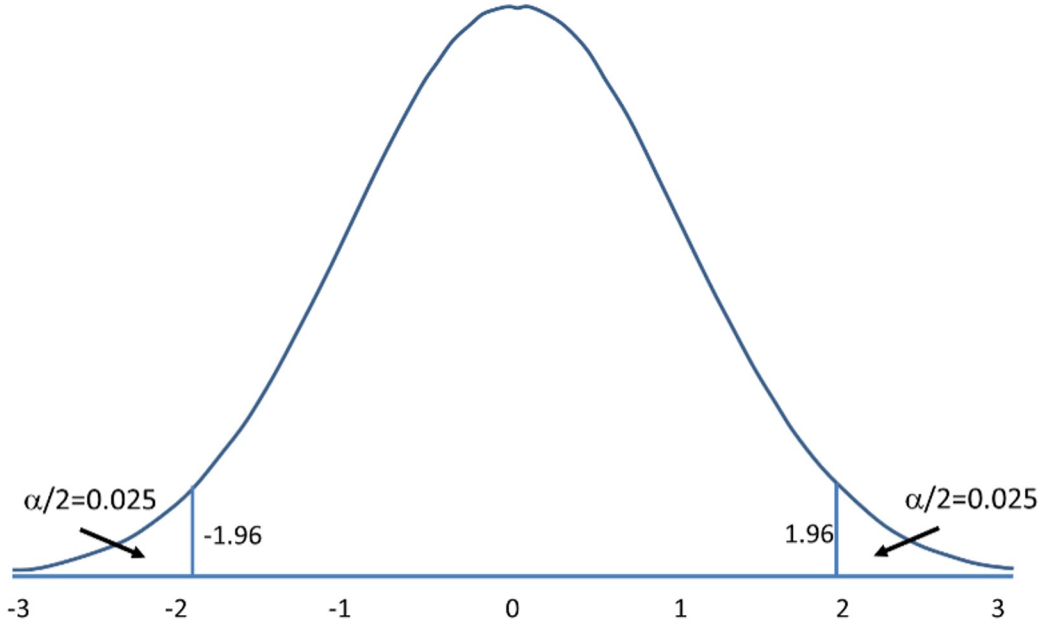
2. p-value computation: under the null-hypothesis $\bar{X} \sim N(\mu_0, \sigma^2/n)$

$$p - value = P(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|) = 2\Phi\left(-\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right)$$

   where $\Phi()$ denotes the cumulative distribution function of the normal

3. decision: given significance level α , we reject H_0 iff $\alpha \geq p - value$

Two tailed test



The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

Two-Tailed Test	
α	Z
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

Extension to large samples

- The results on testing means of normals can be extended to large sample test where the test statistic is approximately (in the asymptotic sense) normally distributed.
- Common examples are given by the z-test (for >30 sample points) and the t-test (to be used with a lower number of samples)

- Example 1 – Testing mean: let $\{X_1, \ldots, X_n\}$ be iid samples from some population distribution with unknown mean μ.
  - One-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu > \mu_0$
  - Two-Sided Test: $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$
- Test statistic is sample mean $\bar{X}$. By central limit theorem $\bar{X} \to N(\mu, \sigma^2/n)$

  All formulae we have obtained previously are valid.
- When σ is unknown, one can use sample standard deviation s in place of σ

Extension to large samples

- The results on testing means of normals can be extended to large sample test where the test statistic is approximately (in the asymptotic sense) normally distributed.
- Common examples are given by the z-test (for >30 sample points) and the t-test (to be used with a lower number of samples)

- Example 2 – Testing proportions: let $\{X_1, \ldots, X_n\}$ be iid Bernoulli samples such that $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$

  - One-Sided Test: $H_0 : p = p_0$, $H_a : p > p_0$
  - Two-Sided Test: $H_0 : p = p_0$, $H_a : p \neq p_0$

- Test statistic is sample mean $\bar{X}$. By central limit $\bar{X} \to N(p, p(1-p)/n)$

  All formulae we have obtained previously are valid with p_0 in place of μ_0 and $p_0(1\text{-}p_0)$ in place of σ^2

Example: one sample z-test

- Suppose we have a sample with: $\bar{x} = 0.52, \sigma = 7.89, n = 27$

$$H_0 : \mu = 0, \qquad H_a : \mu > 0$$

- Compute standard z-test statistic:

$$z = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} = \frac{0.52}{7.89/\sqrt{27}} = 0.3425$$

- Compute p-value: $\Phi(-z) = \Phi(-0.3425 = 0.366)$

- Decision: for $\alpha = 0.05$ we accept H_0 as $0.366 > 0.05$

Example: two sample z-test

- Compare two population means: Do indoor cats live longer than outdoor ones?

Cats	Sample size	Mean age	Sample Std
Indoor	64	14	4
Wild	36	10	5

- State hypotheses: let μ_I (resp., μ_O) denote the true population mean age of indoor (resp., outdoor) cats

$$H_0 : \mu_I = \mu_O, \quad H_a : \mu_I > \mu_O$$

- Test statistic: difference in population means

$$\bar{d} = \bar{x}_I - \bar{x}_O = 14 - 10 = 4$$

Example: two sample z-test

- Characterize distribution $\bar{D} = \bar{X}_I - \bar{X}_O$:

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{4^2}{14} + \frac{5^2}{10}} = 0.97$$

- p-value computation: under the null hypothesis $\bar{D} \sim N(0, \sigma_{\bar{D}}^2)$

$$p - value = P(N(0, 0.97^2) \geq 4) = 1 - \Phi\left(\frac{4}{0.97}\right) \leq 0.00003$$

- Decions: for confidence $\alpha = 0.05$ we reject the null hypothesis

Example: Comparing Two Proportions:

- In order to test if there is any significant difference between opinions of countries A and B on gun ban, random samples of 100 people from country A and 150 people from country B were taken.

Country	Sample size	Favor	Oppose
A	100	52	48
B	150	95	55

- Set up the hypotheses: let p_A (resp., p_B) be the fraction of people from country A (resp., country B) which support gun ban.

$$H_0 : p_A = p_B, \qquad H_a = p_A \neq p_B$$

Example: Comparing Two Proportions:

- Test statistic: difference in sample (empirical) proportions:

$$\overline{D} = \overline{p}_A - \overline{p}_B = \frac{52}{100} - \frac{95}{150} = -0.113$$

- Distribution of difference of sample proportions: $\bar{D}$ is approximately normal with $\mu = p_A - p_B$ and:

$$\overline{\sigma}_D = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

- Pooled estimate: Under the null hypothesis $p_A = p_B$. Hence we can compute a pooled estimate for $p_A = p_B$ as:

$$\bar{p} = \frac{52 + 95}{100 + 150} = 0.588$$

Example: Comparing Two Proportions:

- p-value: Under the null-hypotheis we have $\bar{D} \sim N(0, \sigma^2_{\bar{D}})$, where:

$$\sigma^2_{\bar{D}} = \bar{p}(1-\bar{p})\left(\frac{1}{n_F} + \frac{1}{n_F}\right)$$

$$= 0.588(1-0.588)(100^{-1} + 150^{-1}) = 0.0040373316$$

$\sigma^2_{\bar{D}}$ is the sample variance

Two tailed test, hence

$$p - values = P(|\bar{D} - \mu_0| \geq |\bar{d} - \mu_0|) = 2\Phi\left(-\frac{|\bar{d} - \mu_0|}{\sigma_{\bar{D}}}\right)$$

$$= 2\Phi\left(-\frac{0.113}{0.6354}\right) = 0.075$$

Example: Comparing Two Proportions:

- Decision: given the confidence level $\alpha = 0.05$, we accept the null hypothesis, and, thus we reject the alternative hypothesis.

- There is no statistically significant evidence that suggests people from country A and country B have different opinions on gun ban.

CS145: Lecture 20 Outline

- Frequentist Hypothesis Tests
- Bayesian Hypothesis Tests

Bayesian vs Frequentist approach

- **Frequentist:**
  - *Fixed:* The true (but unknown) state of the hypothesis in the world.
  - *Random:* The data, over many hypothetical repetitions of experiment.

Does the data provide enough evidence to reject a null-hypothesis with confidence?

- **Bayesian:**
  - *Fixed:* The single data set we have observed.
  - *Random:* The true value of the hypothesis, given our partial knowledge.

What is the hypothesis which is most likely to be correct?

Bayesian Hypothesis Testing

Also known as classification, categorization, or discrimination.

We want to choose between two *mutually exclusive hypotheses*:

- *H=0:* *Null* hypothesis
- *H=1:* *Alternative* hypothesis

There is some *prior probability* of each hypothesis:

- Probability of *H=0*: $p_H(0) = q$
- Probability of *H=1*: $p_H(1) = 1 - q$

Observed data *X* has a *likelihood function* under each hypothesis:

- Discrete data: $p_{X|H}(x \mid 0), \quad p_{X|H}(x \mid 1)$
- Continuous data: $f_{X|H}(x \mid 0), \quad f_{X|H}(x \mid 1)$

Formulas on following slides assume discrete X for simplicity.

Posterior Probabilities of Hypotheses

Bayesian hypothesis testing procedures assume that:

- The true value of the hypothesis is a *random variable*
- The *prior distribution* encodes previously observed data.
  If no prior knowledge, set $p_H(0) = p_H(1) = 0.5$
- We have a single new observation $X=x$, with *likelihood*

$$p_{X|H}(x \mid 0), \quad p_{X|H}(x \mid 1)$$

Compute *posterior probability of hypothesis* via Bayes rule:

$$p_{H|X}(h \mid x) = \frac{p_{X|H}(x \mid h)p_H(h)}{p_X(x)} \qquad p_{H|X}(0 \mid x) + p_{H|X}(1 \mid x) = 1$$

$$p_X(x) = p_H(0)p_{X|H}(x \mid 0) + p_H(1)p_{X|H}(x \mid 1)$$

Typically both hypotheses have positive probability. How should we choose?

Loss Functions

We need to formalize the notion of the *cost of a mistake*:

$$L(h, g) = \text{cost of predicting hypothesis } g \text{ when } h \text{ is true.}$$

Properties of standard *loss functions* used for hypothesis testing:

- Assume there is *no loss for correct decisions*:

$$L(0, 0) = L(1, 1) = 0$$

- **Type I Error:** Positive loss for *false positives* or “false alarms”

$$L(0, 1) = \lambda_{01} > 0$$

- **Type II Error:** Positive loss for *false negatives* or “missed detections”

$$L(1, 0) = \lambda_{10} > 0$$

- Can encode “utilities” or “rewards” as negative losses

Example: Spam Classification

$p_{X|H}(x \mid h) =$ *Model of words in email: naïve Bayes, Markov chain, …*

Decision	*h=0*: Ham (not spam)	*h=1*: Spam
$g = 0$	$L(0,0) = 0$	$L(1,0) = \lambda_{10} > 0$ ***False negative:*** *A spam email is placed in your Inbox.*
$g = 1$	$L(0,1) = \lambda_{01} > 0$ ***False positive:*** *Some real email is placed in Spam folder.*	$L(1,1) = 0$

Example: Biometric Identification

$f_{X|H}(x \mid h) =$ *Features from phone's camera, fingerprint sensor, …*

Decision	*h=0*: Authorized unlock	*h=1*: Attacker
$g = 0$	$L(0,0) = 0$	$L(1,0) = \lambda_{10} > 0$ ***False negative:*** *Attacker gains unauthorized access to phone!*
$g = 1$	$L(0,1) = \lambda_{01} > 0$ ***False positive:*** *Enter biometric data again or enter passcode.*	$L(1,1) = 0$

Example: Medical Diagnosis

$f_{X|H}(x \mid h) =$ *Results of various laboratory tests, scans, …*

Decision	*h=0*: Healthy	*h=1*: Serious Illness
$g = 0$	$L(0,0) = 0$	$L(1,0) = \lambda_{10} > 0$ ***False negative:*** *Illness goes untreated and you become more sick.*
$g = 1$	$L(0,1) = \lambda_{01} > 0$ ***False positive:*** *Unnecessary painful or costly medical tests.*	$L(1,1) = 0$

Bayesian Decision Theory

We are given both a *probabilistic model* and a *loss function*:

Posterior distribution:

$$p_{H|X}(h \mid x) = \frac{p_{X|H}(x \mid h)p_H(h)}{p_X(x)}$$

Loss function:

$$L(0,1) = \lambda_{01} > 0 \qquad L(1,0) = \lambda_{10} > 0$$

The optimal decision then *minimizes the posterior expected loss*:

$$\delta(x) = \arg\min_g \; E[L(h,g) \mid X = x] = \arg\min_g \sum_{h=0}^{1} L(h,g)p_{H|X}(h \mid x)$$

Likelihood Ratio Tests

Expected loss of guessing hypothesis *h=1*:

$$L(0,1)p_{H|X}(0 \mid x) + L(1,1)p_{H|X}(1 \mid x) = \lambda_{01}p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis *h=0*:

$$L(0,0)p_{H|X}(0 \mid x) + L(1,0)p_{H|X}(1 \mid x) = \lambda_{10}p_{H|X}(1 \mid x)$$

The optimal decision then *minimizes the posterior expected loss*:

$$\delta(x) = \arg\min_g \; E[L(h,g) \mid X = x] = \arg\min_g \sum_{h=0}^{1} L(h,g)p_{H|X}(h \mid x)$$

Likelihood Ratio Tests

Expected loss of guessing hypothesis *h=1*:

$$L(0,1)p_{H|X}(0 \mid x) + L(1,1)p_{H|X}(1 \mid x) = \lambda_{01}p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis *h=0*:

$$L(0,0)p_{H|X}(0 \mid x) + L(1,0)p_{H|X}(1 \mid x) = \lambda_{10}p_{H|X}(1 \mid x)$$

It is optimal to decide *h=1* if and only if:

$$\lambda_{01}p_{H|X}(0 \mid x) \leq \lambda_{10}p_{H|X}(1 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \cdot \left(\frac{\lambda_{01}}{\lambda_{10}}\right) \qquad p_H(0) = q$$

Minimizing Probability of Error

The general *likelihood ratio test* picks *h=1* if and only if:

$$\lambda_{10} p_{H|X}(1 \mid x) \geq \lambda_{01} p_{H|X}(0 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \cdot \left(\frac{\lambda_{01}}{\lambda_{10}}\right) \qquad p_H(0) = q$$

If *all errors are equally costly* this simplifies: $\lambda_{10} = \lambda_{01} = 1$

$$p_{H|X}(1 \mid x) \geq p_{H|X}(0 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right)$$

Pick hypothesis with larger posterior probability to minimize number of errors

Minimizing Probability of Error

The general *likelihood ratio test* picks *h=1* if and only if:

$$\lambda_{10} p_{H|X}(1 \mid x) \geq \lambda_{01} p_{H|X}(0 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \cdot \left(\frac{\lambda_{01}}{\lambda_{10}}\right) \qquad p_H(0) = q$$

If *all errors are equally costly*, and *hypotheses have equal prior probability*:

$$\lambda_{10} = \lambda_{01} = 1$$
$$q = 0.5$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq 1$$

Pick hypothesis with larger likelihood to minimize number of errors

Bayesian vs Frequentist approach

- **Bayesian:**
  - *Fixed:* The single data set we have observed.
  - *Random:* The true value of the hypothesis, given our partial knowledge.
- **Frequentist:**
  - *Fixed:* The true (but unknown) state of the hypothesis in the world.
  - *Random:* The data, over many hypothetical repetitions of experiment.

- **Bayesian:** Set threshold to *minimize expected loss* $L(h, g)$

$$\xi = \left(\frac{p_H(0)}{p_H(1)}\right) \cdot \left(\frac{L(0,1)}{L(1,0)}\right)$$

- **Frequentist:** Set threshold to *control false positive rate*

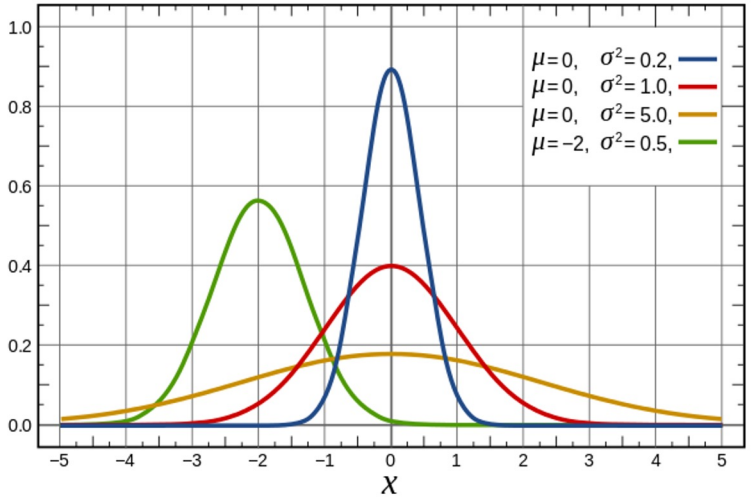
$$P(L(X) > \xi; h = 0) = \alpha$$

Example: Gaussian Hypothesis Tests

$$p_H(0) = q$$

$$f_{X|H}(x \mid 1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$$

$$f_{X|H}(x \mid 0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2}$$



Assuming all errors are equally costly, we choose *h=1* if:

$$\frac{f_{X|H}(x \mid 1)}{f_{X|H}(x \mid 0)} \geq \left(\frac{q}{1-q}\right) \qquad c = \log\left(\frac{q}{1-q}\right)$$

$$\log(f_{X|H}(x \mid 1)) - \log(f_{X|H}(x \mid 0)) \geq c$$

Example: Gaussian Hypothesis Tests

$$\log(f_{X|H}(x \mid i)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2$$

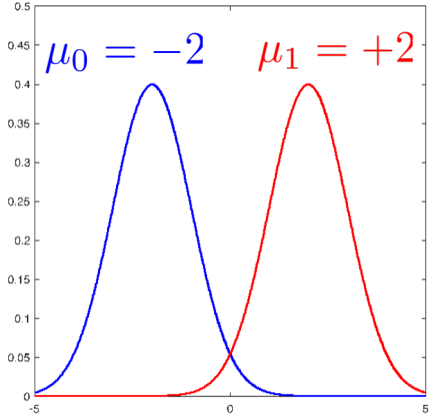
Suppose that $\sigma_1 = \sigma_0 = \sigma$ and $\mu_1 > \mu_0$:

$$\log(f_{X|H}(x \mid 1)) - \log(f_{X|H}(x \mid 0)) \geq c$$

$$-\frac{1}{2\sigma^2}(x-\mu_1)^2 + \frac{1}{2\sigma^2}(x-\mu_0)^2 \geq c$$

With some algebra, we choose *h=1* if:

$$x \geq \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2 c}{\mu_1 - \mu_0}$$



Example: Gaussian Hypothesis Tests

$$\log(f_{X|H}(x \mid i)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2$$

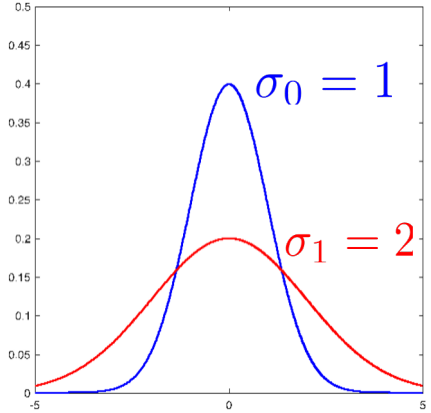
Suppose that $\mu_1 = \mu_0 = 0$ and $\sigma_1 > \sigma_0$:

$$\log(f_{X|H}(x \mid 1)) - \log(f_{X|H}(x \mid 0)) \geq c$$

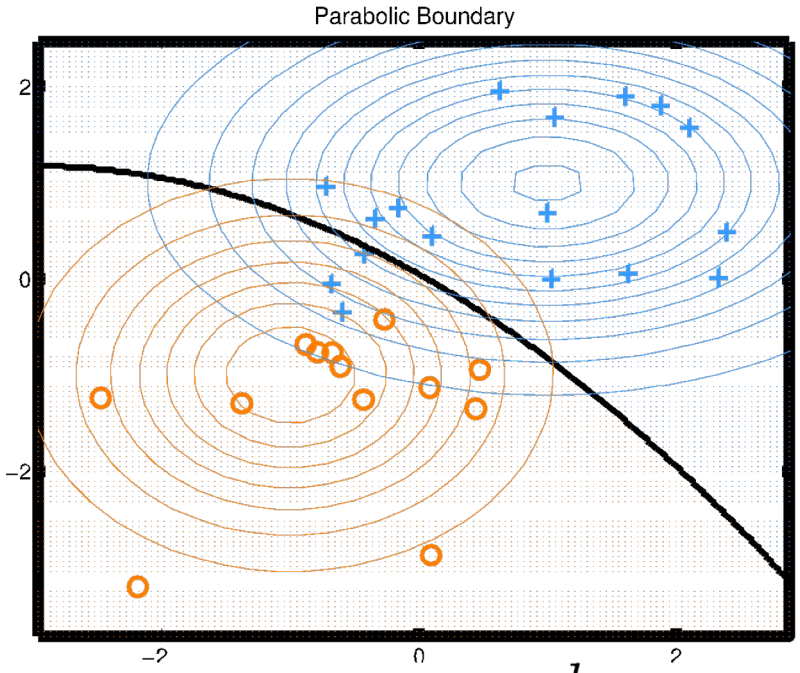
$$-\frac{x^2}{2\sigma_1^2} + \frac{x^2}{2\sigma_0^2} - \frac{1}{2}\log(\sigma_1^2) + \frac{1}{2}\log(\sigma_0^2) \geq c$$

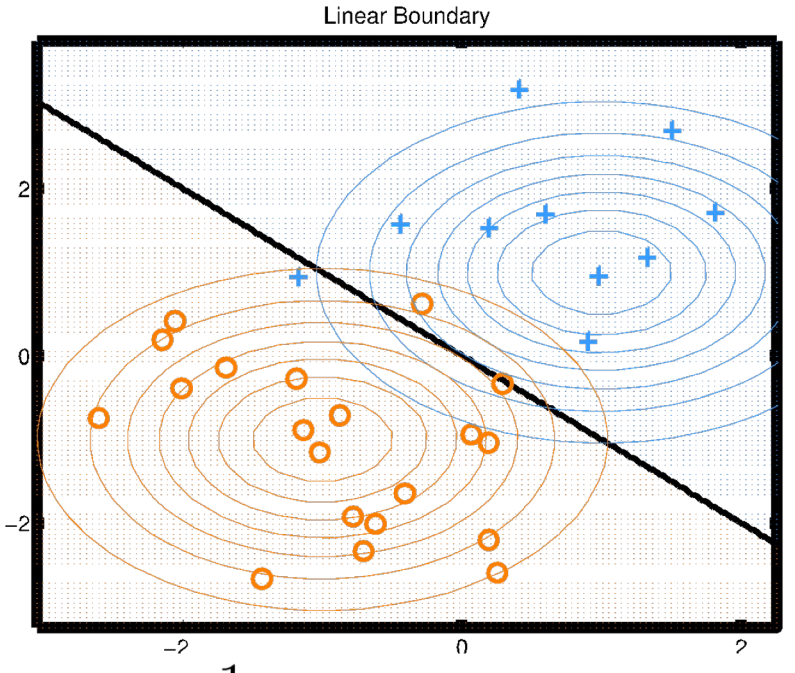
With some algebra, we choose *h=1* if:

$$x^2 \geq \frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2 - \sigma_0^2}\left(c + \log\frac{\sigma_1}{\sigma_0}\right)$$



Multivariate Gaussian Likelihoods





$$\log(f_{X|H}(x \mid i)) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)$$

- Decision boundary is always a *quadratic function*
- If classes have same covariance, decision boundary is a *linear function*