CS145: Lecture 20 Outline

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

\[ \text{HTTHTHTHHTTHTHTHHTTT} \]

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

\[ \text{HTTHTHTHHTTHTHTHHTTT} \]

what is the likelihood that the coin is fair (unbiased)?
Hypothesis testing and coin flips

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.

Question: Is the coin fair? 

\[ \hat{p} = \frac{20217}{40000} \]
The Null and Alternative Hypothesis

$H_0$: The null hypothesis - the default model unless proven differently

$H_1$: The alternative hypothesis – if the observed data is not consistent with the null hypothesis

The defended is innocent ($H_0$) unless proven guilty ($H_1$)

Statistical hypothesis testing rejects or doesn’t reject the null hypothesis – it doesn’t verify the alternative
We need to decide regarding our coin…

- \( H_1: \) The alternative hypothesis - the coin is biased
- \( H_0: \) 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 Idea in Frequentist Test

- Hypotheses are **fixed**: representing two possible models for the process that generated the data.

- Data is **random sample**: the analyst evaluates the likelihood of the observed data in the two models.
The hypothesis testing set-up

1. Set Up *Null Hypothesis* $(H_0)$ and *Alternative Hypothesis* $(H_a)$:

$$H_0 : p = 0.5 \quad H_a : p \neq 0.5$$

Null hypothesis is the claim to be tested.

Hypothesis testing evaluates the strength of empirical evidence against null hypothesis.

2. Finding a **Test Statistic**: here the empirical frequency \( \hat{p} \).

3. Determine If Data Is **Plausible**, assuming null hypothesis is correct
The hypothesis testing set-up

1. Set Up Null Hypothesis ($H_0$) and Alternative Hypothesis ($H_a$):

$$H_0 : p = 0.5 \quad H_a : p \neq 0.5$$

Null hypothesis is the claim to be tested.

Hypothesis testing evaluates the strength of empirical evidence against null hypothesis.

2. Finding a Test Statistic: What else but empirical frequency $\hat{p}$.

3. Determine If Data Is Plausible, assuming null hypothesis is correct .... i.e. assuming the population distribution corresponds to $H_0$.
Hypothesis testing with p-value

What Is 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.

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

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

Find p-value: Under null hypothesis $H_0$, $\hat{p}$ is approximately $N(0.5, 0.0025)$. Answer: p-value = 0.03

P-value computation depends on the specific assumptions/test being used!
What Is Significance (Confidence) Level $\alpha$?
It is an artificially given threshold such that any empirical observation that falls in the top $\alpha$ proportion of the most extreme scenarios (under null hypothesis $H_0$) is deemed implausible. The most common value of $\alpha$ is 5%, even though $\alpha = 1\%$ is also widely used.

Null hypothesis is rejected if and only if the corresponding P-value is less than the significance level $\alpha$. 
Why is this working?

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

Hence, $P(p\text{-value } H_0 \leq \alpha) = \alpha$

Thus, $P(H_0 \text{ rejected } | H_0 \text{ is a true null }) = \alpha$
Relation Between P-Value and Significance Level $\alpha$: Null hypothesis is rejected if and only if P-value is less than the significance level $\alpha$.

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

<table>
<thead>
<tr>
<th>Decision</th>
<th>Actual Situation</th>
<th>Outcome (Probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Do Not Reject $H_0$</strong></td>
<td>$H_0$ True</td>
<td>No error $(1 - \alpha)$</td>
</tr>
<tr>
<td></td>
<td>$H_0$ False</td>
<td>Type II Error $(\beta)$</td>
</tr>
<tr>
<td><strong>Reject $H_0$</strong></td>
<td></td>
<td>Type I Error $(\alpha)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No Error $(1 - \beta)$</td>
</tr>
</tbody>
</table>
Types of error and Power of Test

- **Type I Error (False Positive):** Given a significance level $\alpha$, 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 $\alpha$, 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 $\alpha$ and $\beta$ 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.
Let \( \{X_1, \ldots, X_n\} \) be iid samples from \( \text{N}(\mu, \sigma^2) \), where \( \sigma^2 \) is known but \( \mu \) unknown. Want to perform hypothesis testing on \( \mu \).

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**: \( H_0 : \mu = \mu_0, \ H_a : \mu \neq \mu_0 \)
Let \( \{X_1, \ldots, X_n\} \) be iid samples from \( \text{N}(\mu, \sigma^2) \), where \( \sigma^2 \) is known but \( \mu \) unknown. Want to perform hypothesis testing on \( \mu \).

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 $\alpha$, we reject $H_0$ iff $\alpha \geq p - value$
The decision rule is: Reject $H_0$ if $Z \geq 1.645$. 

<table>
<thead>
<tr>
<th>Upper-Tailed Test</th>
<th>$\alpha$</th>
<th>$Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.10</td>
<td>1.282</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1.645</td>
</tr>
<tr>
<td></td>
<td>0.025</td>
<td>1.960</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>2.326</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>2.576</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>3.090</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>3.719</td>
</tr>
</tbody>
</table>
Use the appropriate table!

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

Report the cumulative area from the LEFT

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

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

**NEGATIVE z Scores**

<table>
<thead>
<tr>
<th></th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0</td>
<td>.4900</td>
<td>.4959</td>
<td>.4998</td>
<td>.5037</td>
<td>.5076</td>
<td>.5114</td>
<td>.5152</td>
<td>.5189</td>
<td>.5227</td>
<td>.5263</td>
</tr>
<tr>
<td>-0.1</td>
<td>.4821</td>
<td>.4861</td>
<td>.4893</td>
<td>.4924</td>
<td>.4954</td>
<td>.4983</td>
<td>.5011</td>
<td>.5038</td>
<td>.5065</td>
<td>.5092</td>
</tr>
<tr>
<td>-0.2</td>
<td>.4740</td>
<td>.4774</td>
<td>.4807</td>
<td>.4839</td>
<td>.4870</td>
<td>.4901</td>
<td>.4932</td>
<td>.4961</td>
<td>.4990</td>
<td>.5018</td>
</tr>
<tr>
<td>-0.3</td>
<td>.4655</td>
<td>.4685</td>
<td>.4714</td>
<td>.4743</td>
<td>.4771</td>
<td>.4798</td>
<td>.4825</td>
<td>.4851</td>
<td>.4876</td>
<td>.4900</td>
</tr>
<tr>
<td>-0.4</td>
<td>.4565</td>
<td>.4591</td>
<td>.4616</td>
<td>.4640</td>
<td>.4663</td>
<td>.4685</td>
<td>.4707</td>
<td>.4728</td>
<td>.4748</td>
<td>.4767</td>
</tr>
<tr>
<td>-0.5</td>
<td>.4465</td>
<td>.4488</td>
<td>.4510</td>
<td>.4531</td>
<td>.4552</td>
<td>.4572</td>
<td>.4591</td>
<td>.4609</td>
<td>.4626</td>
<td>.4643</td>
</tr>
<tr>
<td>-0.6</td>
<td>.4355</td>
<td>.4375</td>
<td>.4394</td>
<td>.4412</td>
<td>.4430</td>
<td>.4447</td>
<td>.4464</td>
<td>.4480</td>
<td>.4495</td>
<td>.4510</td>
</tr>
<tr>
<td>-0.7</td>
<td>.4243</td>
<td>.4259</td>
<td>.4274</td>
<td>.4289</td>
<td>.4303</td>
<td>.4317</td>
<td>.4330</td>
<td>.4343</td>
<td>.4355</td>
<td>.4367</td>
</tr>
<tr>
<td>-0.8</td>
<td>.4129</td>
<td>.4142</td>
<td>.4154</td>
<td>.4166</td>
<td>.4177</td>
<td>.4188</td>
<td>.4198</td>
<td>.4208</td>
<td>.4217</td>
<td>.4226</td>
</tr>
<tr>
<td>-0.9</td>
<td>.3997</td>
<td>.4005</td>
<td>.4012</td>
<td>.4018</td>
<td>.4024</td>
<td>.4029</td>
<td>.4034</td>
<td>.4039</td>
<td>.4043</td>
<td>.4047</td>
</tr>
<tr>
<td>-1.0</td>
<td>.3699</td>
<td>.3704</td>
<td>.3708</td>
<td>.3712</td>
<td>.3715</td>
<td>.3718</td>
<td>.3721</td>
<td>.3723</td>
<td>.3725</td>
<td>.3727</td>
</tr>
</tbody>
</table>
Example

Suppose we need $\Phi(-2.45)$.
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 $\Phi(\cdot)$ denotes the cumulative distribution function of the normal.

3. **decision**: given significance level $\alpha$, we reject $H_0$ iff $\alpha \geq p-value$
The decision rule is: Reject $H_0$ if $Z \leq 1.645$. 

<table>
<thead>
<tr>
<th>Lower-Tailed Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$Z$</td>
</tr>
<tr>
<td>0.10</td>
<td>-1.282</td>
</tr>
<tr>
<td>0.05</td>
<td>-1.645</td>
</tr>
<tr>
<td>0.025</td>
<td>-1.960</td>
</tr>
<tr>
<td>0.010</td>
<td>-2.326</td>
</tr>
<tr>
<td>0.005</td>
<td>-2.576</td>
</tr>
<tr>
<td>0.001</td>
<td>-3.090</td>
</tr>
<tr>
<td>0.0001</td>
<td>-3.719</td>
</tr>
</tbody>
</table>
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 $\alpha$, we reject $H_0$ iff $\alpha \geq p-value$
The decision rule is: Reject $H_0$ if $Z \leq -1.960$ or if $Z \geq 1.960$. 

### Two-Tailed Test

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>1.282</td>
</tr>
<tr>
<td>0.10</td>
<td>1.645</td>
</tr>
<tr>
<td>0.05</td>
<td>1.960</td>
</tr>
<tr>
<td>0.010</td>
<td>2.576</td>
</tr>
<tr>
<td>0.001</td>
<td>3.291</td>
</tr>
<tr>
<td>0.0001</td>
<td>3.819</td>
</tr>
</tbody>
</table>
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 \( \mu \).
  
  - 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 \( \bar{X} \rightarrow N(\mu, \sigma^2/n) \)
  
  All formulae we have obtained previously are valid.

- When \( \sigma \) is unknown, one can use sample standard deviation \( s \) in place of \( \sigma \).
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 \( \mu_0 \) and \( p_0(1-p_0) \) in place of \( \sigma^2 \)
Example: one sample z-test

- Suppose we have a sample with: \( \bar{x} = 0.52, \sigma = 7.89, n = 27 \)
  
  \[ H_0 : \mu = 0, \quad H_a : \mu > 0 \]

- 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?

<table>
<thead>
<tr>
<th>Cats</th>
<th>Sample size</th>
<th>Mean age</th>
<th>Sample Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor</td>
<td>64</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Wild</td>
<td>36</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

- **State hypotheses:** let $\mu_i$ (resp., $\mu_o$) 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 \( \bar{D} = \bar{X}_I - \bar{X}_O \):
  \[
  \sigma_{\bar{D}} = \sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}} \approx \sqrt{\frac{4^2}{14} + \frac{5^2}{10}} = 0.97
  \]

- p-value computation: under the null hypothesis \( \bar{D} \sim N(0, \sigma^2_{\bar{D}}) \)
  \[
  p-value = P(N(0, 0.97^2) \geq 4) = 1 - \Phi \left( \frac{4}{0.97} \right) \leq 0.00003
  \]

- Decisions: 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 males and females on gun ban, random samples of 100 males and 150 females were taken.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Sample size</th>
<th>Favor</th>
<th>Oppose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>100</td>
<td>52</td>
<td>48</td>
</tr>
<tr>
<td>Female</td>
<td>150</td>
<td>95</td>
<td>55</td>
</tr>
</tbody>
</table>

- Set up the hypotheses: let $p_F$ (resp., $p_M$) be the fraction of females (resp., males) which support gun ban.

$$H_0 : p_F = p_M, \quad H_a : p_F \neq p_M$$
Example: Comparing Two Proportions:

- **Test statistic**: difference in sample (empirical) proportions:
  
  \[ \bar{d} = \bar{p}_F - \bar{p}_M = \frac{52}{100} - \frac{95}{150} = -0.113 \]

- **Distribution of difference of sample proportions**: \( \bar{D} \) is approximately normal with \( \mu = \bar{p}_F - \bar{p}_M \) and:
  
  \[ \sigma_{\bar{D}} = \sqrt{\frac{p_F(1-p_F)}{n_F} + \frac{p_M(1-p_M)}{n_M}} \]

- **Pooled estimate**: Under the null hypothesis \( p_F = p_M \). Hence we can compute a pooled estimate for \( p_F = p_M \) as:
  
  \[ \bar{p} = \frac{52 + 95}{100 + 150} = 0.588 \]
Example: Comparing Two Proportions:

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

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

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

  $\sigma^2_D$ is the sample variance

- Two tailed test, hence

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

  $$= 2\Phi \left( -\frac{0.113}{0.6354} \right) = 0.075$$
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 male and females have different opinions on gun ban.
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: \( p_{X|H}(x \mid 0) \), \( p_{X|H}(x \mid 1) \)
- Continuous data: \( f_{X|H}(x \mid 0) \), \( f_{X|H}(x \mid 1) \)

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* \( p_{X|H}(x|0), p_{X|H}(x|1) \)

Compute *posterior probability of hypothesis* via Bayes rule:

\[
p_{H|X}(h|x) = \frac{p_{X|H}(x|h)p_H(h)}{p_X(x)}
\]

\[
p_{H|X}(0|x) + p_{H|X}(1|x) = 1
\]

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

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

<table>
<thead>
<tr>
<th>Decision</th>
</tr>
</thead>
</table>
| \( g = 0 \) | \( h = 0: \) Ham (not spam) \( L(0, 0) = 0 \) \( L(1, 0) = \lambda_{10} > 0 \)  
|          | \( h = 1: \) Spam \( L(0, 1) = \lambda_{01} > 0 \) \( L(1, 1) = 0 \)  
| \( g = 1 \) | \( \)  

*False positive:*  
Some real email is placed in Spam folder.

*False negative:*  
A spam email is placed in your Inbox.
### Example: Biometric Identification

\[ f_{X|H}(x \mid h) = \text{Features from phone’s camera, fingerprint sensor, …} \]

<table>
<thead>
<tr>
<th>Decision</th>
<th>( h=0 ): Authorized unlock</th>
<th>( h=1 ): Attacker</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g = 0 )</td>
<td>( L(0, 0) = 0 )</td>
<td>( L(1, 0) = \lambda_{10} &gt; 0 )</td>
</tr>
<tr>
<td>( g = 1 )</td>
<td>( L(0, 1) = \lambda_{01} &gt; 0 )</td>
<td>( L(1, 1) = 0 )</td>
</tr>
</tbody>
</table>

- **False positive:** Enter biometric data again or enter passcode.
- **False negative:** Attacker gains unauthorized access to phone!
### Example: Medical Diagnosis

\[ f_{X \mid H}(x \mid h) = \text{Results of various laboratory tests, scans, ...} \]

<table>
<thead>
<tr>
<th>Decision</th>
<th>( h=0 ): Healthy</th>
<th>( h=1 ): Serious Illness</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g = 0 )</td>
<td>( L(0, 0) = 0 )</td>
<td>( L(1, 0) = \lambda_{10} &gt; 0 )</td>
</tr>
<tr>
<td>( g = 1 )</td>
<td>( L(0, 1) = \lambda_{01} &gt; 0 )</td>
<td>( L(1, 1) = 0 )</td>
</tr>
</tbody>
</table>

**False positive:**
Unnecessary painful or costly medical tests.

**False negative:**
Illness goes untreated and you become more sick.
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:**

\[ L(0, 1) = \lambda_{01} > 0 \quad L(1, 0) = \lambda_{10} > 0 \]

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

\[ \delta(x) = \arg \min \limits_g E[L(h, g) \mid X = x] = \arg \min \limits_g \sum_{h=0}^{1} L(h, g)p_{H|X}(h \mid x) \]
Expected loss of guessing hypothesis $h=1$:
\[ L(0, 1)p_{H|X}(0 \mid x) + L(1, 1)p_{H|X}(1 \mid x) = \lambda_1 p_{H|X}(0 \mid x) \]

Expected loss of guessing hypothesis $h=0$:
\[ L(0, 0)p_{H|X}(0 \mid x) + L(1, 0)p_{H|X}(1 \mid x) = \lambda_0 p_{H|X}(1 \mid x) \]

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 $h=1$:

$$L(0, 1)p_{H|X}(0 \mid x) + L(1, 1)p_{H|X}(1 \mid x) = \lambda_0 p_{H|X}(0 \mid x)$$

Expected loss of guessing hypothesis $h=0$:

$$L(0, 0)p_{H|X}(0 \mid x) + L(1, 0)p_{H|X}(1 \mid x) = \lambda_1 p_{H|X}(1 \mid x)$$

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

$$\lambda_0 p_{H|X}(0 \mid x) \leq \lambda_1 p_{H|X}(1 \mid x)$$

$$\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left( \frac{q}{1-q} \right) \cdot \left( \frac{\lambda_0}{\lambda_1} \right)$$

$$p_{H}(0) = q$$
Minimizing Probability of Error

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

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

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

$p_{H}(0) = q$

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 \textit{likelihood ratio test} picks $h=1$ if and only if:

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

\[
\frac{p_{X|H}(x \mid 1)}{p_{X|H}(x \mid 0)} \geq \left( \frac{q}{1-q} \right) \cdot \left( \frac{\lambda_{01}}{\lambda_{10}} \right)
\]

$p_{H}(0) = q$

If \textit{all errors are equally costly}, and \textit{hypotheses have equal prior probability}:

\[
\lambda_{10} = \lambda_{01} = 1
\]

$q = 0.5$

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:

\[
\frac{f_X|_H(x \mid 1)}{f_X|_H(x \mid 0)} \geq \left( \frac{q}{1-q} \right) \]

\[
c = \log \left( \frac{q}{1-q} \right) \]

\[
\log(f_X|_H(x \mid 1)) - \log(f_X|_H(x \mid 0)) \geq c \]
Example: Gaussian Hypothesis Tests

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

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

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

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

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 \( \sigma_1 > \sigma_0 \):

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

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

With some algebra, we choose \( h=1 \) if:

\[ x^2 \geq \frac{2\sigma_1^2 \sigma_0^2}{\sigma_1^2 - \sigma_0^2} \left( c + \log \left( \frac{\sigma_1}{\sigma_0} \right) \right) \]
\[
\log(f_{X|H}(x | i)) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)
\]

- Decision boundary is always a **quadratic function**
- If classes have same covariance, decision boundary is a **linear function**
Report Testing Result Via Rejection/Acceptance and Significance Level:

1) What Is Significance Level $\alpha$?

It is an artificially given threshold such that $\alpha$ any empirical observation that falls in the top $\alpha$ proportion of the most extreme scenarios (under null hypothesis $H_0$) is deemed implausible. The most common value of $\alpha$ is 5%, even though $\alpha = 1\%$ is also widely used.

2) Given Threshold $\alpha$, What Is Rejection Region (RR)?

In this case, the rejection region are extreme scenarios of form

$$RR = \left\{ \left| \hat{p} - \frac{1}{2} \right| \geq c \right\},$$

Where $c$ is chosen such that, under $H_0$ we have $P(RR) = \alpha$.
Testing through coin tosses

- Report Testing Result Via Rejection/Acceptance and Significance Level:
  3) **Determine Rejection Region**

  The rejection region (i.e., the value of $c$) depends on the assumptions regarding the null-hypothesis!

  In our setting

  2) Given Threshold $\alpha$, **What Is Rejection Region (RR)?**

  In this case, the rejection region are extreme scenarios of form

  \[
  RR = \left\{ \left| \hat{p} - \frac{1}{2} \right| \geq c \right\},
  \]

  Where $c$ is chosen such that, under $H_0$ we have $P(RR) = \alpha$