Probability Theory & Statistics

Konstantin Tretyakov (kt@ut.ee)

MTAT.03.239 Bioinformatics 13.10.2010

Keywords

- Probability distribution
- Bayes rule
- MLE, MAP
- Null-model, P-value, T-test, Randomization
- Bonferroni, FDR

Probability Theory Lets You:

- **Formulate** mental models of reality
- **Reason** about those models

Non-probabilistic Model

- Jaak's bike **is black**

- Jaak's bike **has two wheels**

Probabilistic Model

- Jaak's bike colour:
  - P(colour=**black**) = 0.7
  - P(colour=**white**) = 0.3
- Jaak's bike wheels:
  - P(wheels=**2**) = 0.6
  - P(wheels=**3**) = 0.2
  - P(wheels=**1**) = 0.2

Complete Probabilistic Model

P(colour, wheels) =

	wheels=1	wheels=2	wheels=3
colour=black	0.04	0.52	0.14
colour=white	0.16	0.08	0.06

Probabilistic Reasoning

- You want to **pretend** you're Jaak. What bike should you pick?
- You saw Jaak on a 1-wheeled bike. **What was its colour**?
- You observe a 2-wheeled white bike. **Is it Jaak's**?

	wheels=1	wheels=2	wheels=3
colour=black	0.04	0.52	0.14
colour=white	0.16	0.08	0.06

Model Inference (Learning)

- How do you build a **model** of Jaak's bike from **observations**?

Mon:	white,	2 wheels
Tue:	white,	1 wheel
Wed:	black,	3 wheels
Thu:	black,	2 wheels
...		

Statistical Testing

- How do you **answer questions** about Jaak's bike from observations only?

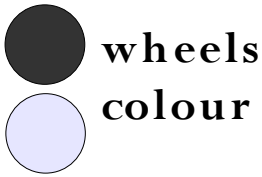
Is bike's colour *independent* of its wheel count?

Probability & Statistics

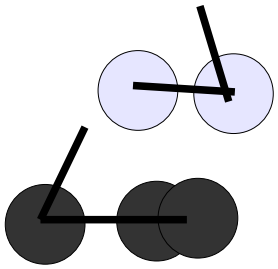
Reality

	wheels=1	wheels=2	wheels=3
colour=black	0.04	0.52	0.14
colour=white	0.16	0.08	0.06

Probabilistic Model



Data Model



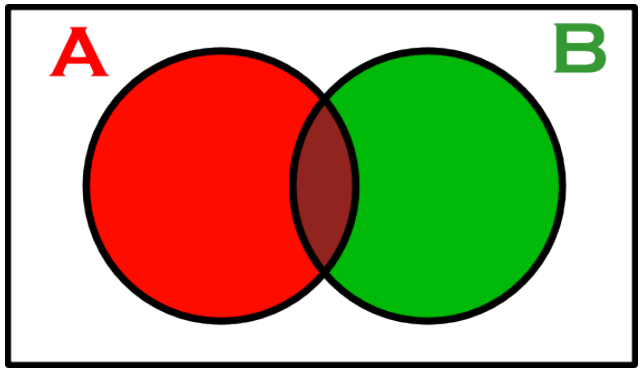
Observations

Probability & Statistics

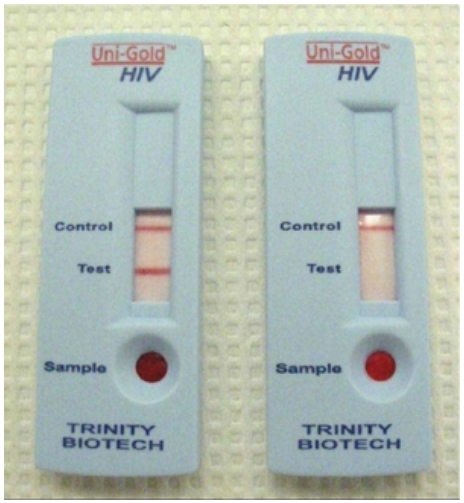
- **Probability Theory**
  - Probability calculus
  - Standard distributions
- **Statistical Theory**
  - Parameter estimation (Model inference)
  - Hypothesis testing
  - Decision making

Probability Theory

- Probability Calculus
  - P(A or B) = P(A) + P(B) – P(A and B)
  - P(A given B) = P(A|B) = P(A and B)/P(B)
  - P(A and B) = P(B) P(A | B)
  - P(A | B) = P(B | A) P(A)/P(B)



Example



$$P(\text{Test Positive} \mid \text{HIV}) = 99\%$$

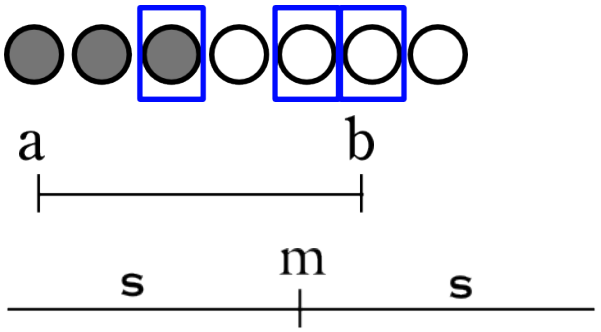
$$P(\text{Test Positive} \mid \sim\text{HIV}) = 1\%$$

$$P(\text{HIV}) = 0.1\%$$

$P(\text{HIV} \mid \text{Test Positive}) = ?$

Probability Theory

- Common Univariate Distributions
  - Bernoulli Be(p)
  - Binomial B(p,n)
  - Poisson P(λ)
  - Hypergeometric HG(N,M,n)
  - Uniform U(a,b)
  - Normal N(m, s^2)



Guess-a-model

- Gene expression measurement
- Presence of an interaction between proteins
- Number of motif matches in a given promoter sequence
- Number of miRNA targets in a gene cluster
- Number of new patients over a year

Probability & Statistics

- **Probability Theory**
  - Probability calculus
  - Standard distributions
- **Statistical Theory**
  - Parameter estimation (Model inference)
  - Hypothesis testing
  - Decision making

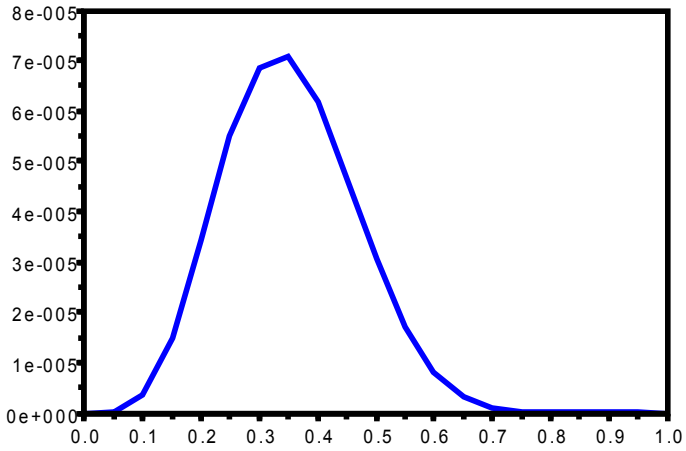
Parameter Estimation

- Let Y ~ Be(**p**)
- Observe y values:
  y = 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0
- What is **p**?

Parameter Estimation

- What is **p**?
  - **Maximum Likelihood Estimation:**

$$L(p) = P(\text{Data} \mid \mathbf{p}) = p^5(1-p)^{10}$$



Parameter Estimation

- What is **p**?
  - **Maximum Likelihood Estimation:**

$$L(p) = P(\text{Data}|\mathbf{p}) = p^5(1-p)^{10}$$

$$l(p) = \log(L(p)) = 5\log(p) + 10\log(1-p)$$

$$u(p) = 5/p - 10/(1-p) = 0$$

$$p=5/15$$

Parameter Estimation

- Let $Y \sim Be(\mathbf{p}\, x_1 + \mathbf{q}\, x_2 + \mathbf{r})$
- Observe (x_1, x_2, y) values:

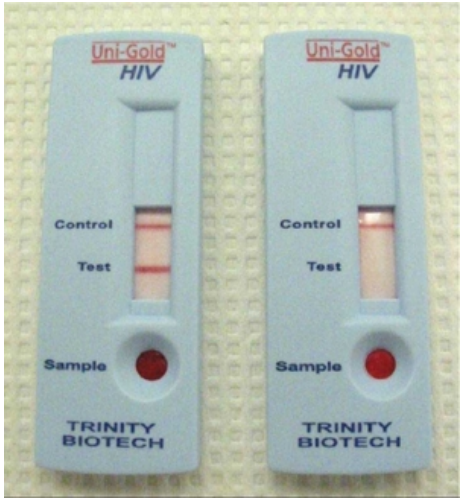
$x_1 = 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 0$

$x_2 = 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0$

$y = 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 0$

- What are **p**, **q**, **r**?

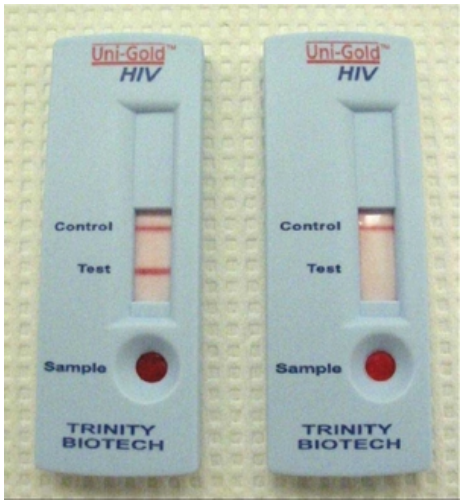
Parameter Estimation



Model parameter: HIV={0,1}

Observation (data): Test Result

Parameter Estimation



Model parameter: HIV={0,1}

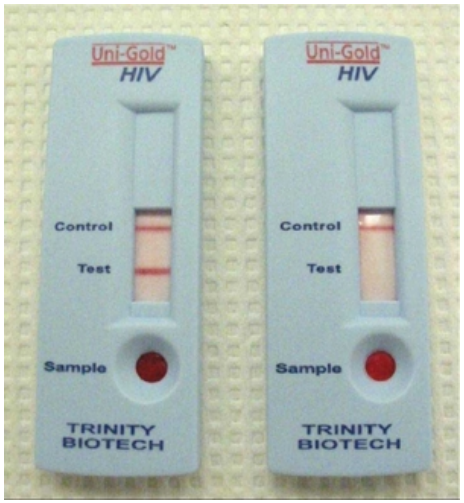
Observation (data): Test Result

Observed: Test positive

$L(1) = P(\text{Test+} \mid \text{HIV}) = 99\%$

$L(0) = P(\text{Test+} \mid \sim\text{HIV}) = 1\%$

Parameter Estimation



Model parameter: HIV={0,1}

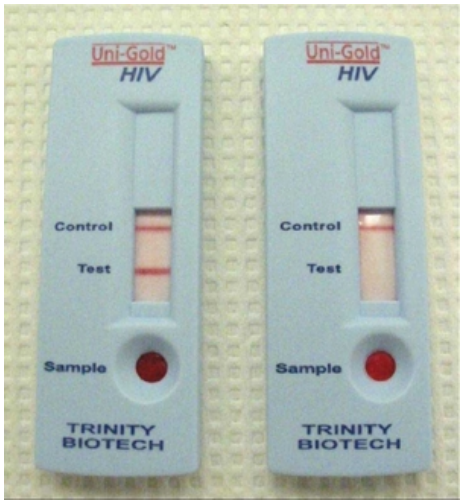
Observation (data): Test Result

Maximum A-posteriori Probability (MAP)-Estimation:

$$P(\text{HIV}\,|\,\text{Test+}) = 0.09$$

$$P(\sim\text{HIV}\,|\,\text{Test+}) = 0.91$$

Parameter Estimation



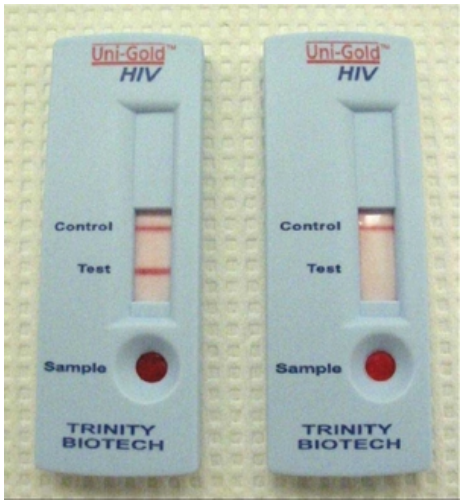
Model parameter: HIV={0,1}

Observation (data): Test Result

Bayesian Estimation:

Result =
{HIV=0.09, ~HIV=0.91}

Parameter Estimation



Model parameter: HIV={0,1}

Observation (data): Test Result

Bayesian Estimation II:

Result =

0.09*1 + 0.91*0 = 0.09

Parameter Estimation

- Task: given model F(**p**) and data X, find **the most appropriate** parameter values **p**
- Methods:
  - **MLE**: p = argmax P(X|p)
  - **MAP**: p = argmax P(p|X)
  - **Bayesian:** p = E(p|X)
  - Other: (e.g. robust or sparse):
    p = argmax f(p,X)

Parameter Estimate Quality

- Confidence intervals

  $n/N - 1/\sqrt{N} < \mathbf{p} < n/N + 1/\sqrt{N}$

- Bayesian confidence intervals

$$q_{2.5} < \mathbf{p} < q_{97.5}$$

- Bootstrap estimates

$$q_{2.5} < \mathbf{p} < q_{97.5}$$

Probability & Statistics

- **Probability Theory**
  - Probability calculus
  - Standard distributions
- **Statistical Theory**
  - Parameter estimation (Model inference)
  - Hypothesis testing
  - Decision making

Hypothesis Testing

- You observe a 2-wheeled white bike. **Is it Jaak's?**

	wheels=1	wheels=2	wheels=3
colour=black	0.04	0.52	0.14
colour=white	0.16	0.08	0.06

Hypothesis Testing

- You observe a 2-wheeled white bike. **Is it Jaak's?**

P-value = 0.08

Significance threshold = 0.10

	wheels=1	wheels=2	wheels=3
colour=black	0.04	0.52	0.14
colour=white	0.16	0.08	0.06

Hypothesis Testing

- **Null-hypothesis:**
  - "The bike belongs to Jaak"
- **Test statistic:** (colour, wheels)
- **P-value:**
  - Probability of observations under null hypothesis
- **Decision rule:**
  - *Reject* null hypothesis if
    p-value < significance threshold

Hypothesis Testing

- Are the expressions of Gene A and Gene B different?

Hypothesis Testing

- Are the expressions of Gene 1 and Gene 2 different?

- **Null hypothesis:**

$$X_1, X_2 \sim N(m, s^2)$$

- **Test statistic:**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{2}{n}}}$$

- **P-value:** $P(T >= t)$

Hypothesis Testing

- Is GCG a significant motif in:
  CCAC**GCG**ATTGAC**GCG**AGCAC**GCG**AGGAAGTACC
  CG**GCG**CGTAAGA**GCG**CAGTCCATTGA**GCG**C**GCG**C

Hypothesis Testing

- Is GCG a significant motif in:
  CCAC**GCG**ATTGAC**GCG**AGCAC**GCG**AGGAAGTACC
  CG**GCG**CGTAAGA**GCG**CAGTCCATTGA**GCG**C**GCG**C
- **Null hypothesis**:
  The string is random
- **Test statistic**:
  #matches = 7
- **P-value**: P(#matches >= 7)

Hypothesis Testing

- Here's another string:
  T**ATC**CTTAACGGTGCCAAG**ATC**CTA**ATCATC**TACA
  G**ATC**TTA**ATC**AGTTGC**ATC**AGCAGGTACTGCT

- We observe that ATC is present 7 times in it. Is it significant?

Multiple Testing

- In the previous example we have implicitly performed 64 tests
- If each of them fails with probability 5%, how many are expected to fail?
- How to deal with it?

Multiple Testing

- Multiple testing corrections:
  - Bonferroni correction
  - Sidak correction
  - False Discovery Rate

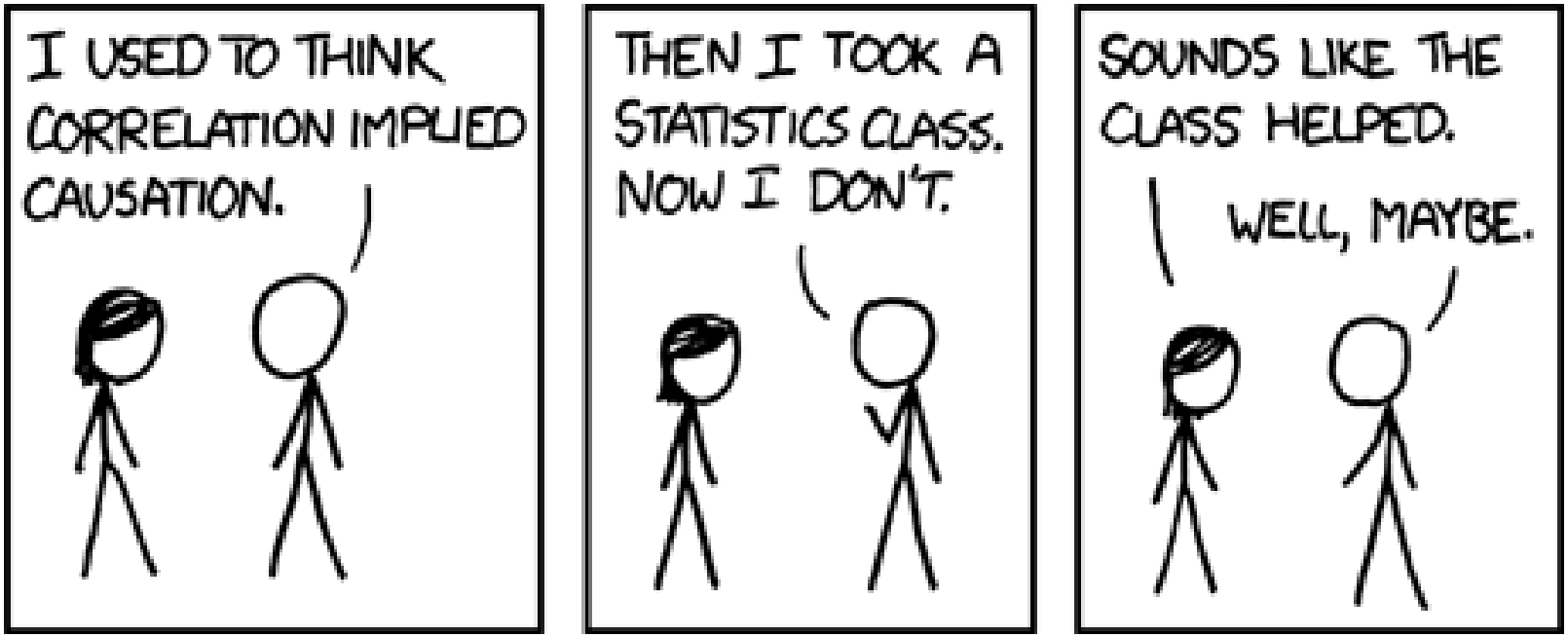
Probability & Statistics

- **Probability Theory**
  - Probability calculus
  - Standard distributions
- **Statistical Theory**
  - Parameter estimation (Model inference)
  - Hypothesis testing
  - Decision making

Keywords

- Probability distribution
- Bayes rule
- MLE, MAP
- Null-model, P-value, T-test, Randomization
- Bonferroni, FDR

Questions


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.