



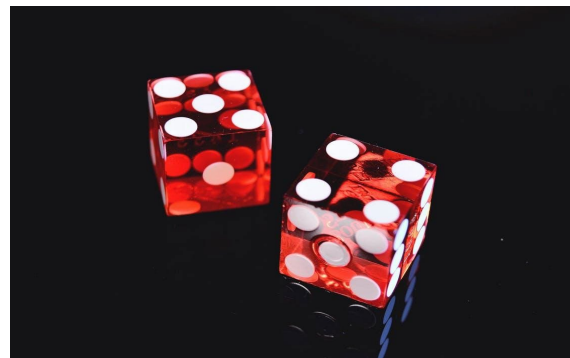
CS 540 Introduction to Artificial Intelligence

Probability & Statistics

University of Wisconsin-Madison
Fall 2025 Sections 1 & 2

Outline

- Probability
 - Basics: definitions and axioms
 - Random Variables (RVs) and joint distributions
 - Independence, conditional probability, chain rule
 - Bayes' Rule and Inference
- Statistics

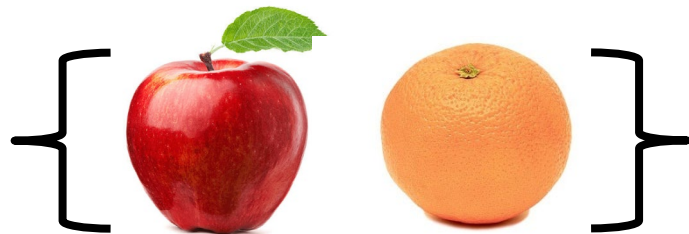


Basics: Random Variables

- Intuitively: a number X that's random
- Mathematically: map random outcomes to real values

$$X : \Omega \rightarrow \mathbb{R}$$

- Why?
 - Previously, everything is a set.
 - Real values are easier to work with



Basics: CDF & PDF

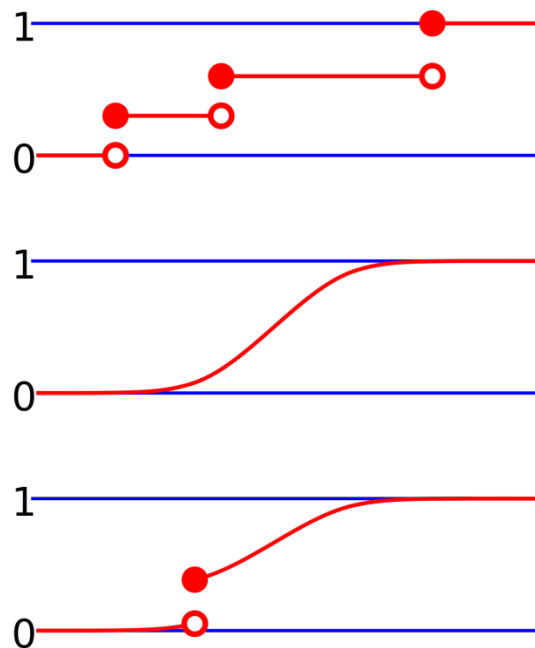
- Can still work with probabilities:

$$P(X = 3)$$

- Cumulative Distribution Func. (CDF)

$$F_X(x) := P(X \leq x)$$

- Density / mass function $p_X(x)$



Wikipedia CDF

Basics: **Expectation & Variance**

- Another advantage of RVs are “summaries”
- Expectation: $E[X] = \sum_a a \times P(x = a)$
 - The “average”
- Variance: $Var[X] = E[(X - E[X])^2]$
 - A measure of “spread”

Basics: Joint Distributions

- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
 - Why? Work with **multiple** types of uncertainty that correlate with each other



Basics: Marginal Probability

- Given a joint distribution $P(X = a, Y = b)$

- Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

- This is the “marginal” distribution.

Date	Item	Amount	Total
1832			
Oct 1	Supper Beer	6	6
5	Supper of Housewife	16	22
"	Breakfast 8 1/2	3	25
Dec 11	Dinner at Club	2 6	27
"	Coffee	6	33
12	Breakfast	1 6	34
13	Breakfast	1 6	36
"	Tea	6	42
14	Breakfast	1 6	43
15	Breakfast	1 6	45
1833			
Jan 20	Tea at dinner club	6	51
27	Breakfast	1 6	52
"	Supper	1	53
Feb 19	Soda Water	6	59
23	Oranges	1 6	60
March 22	Supper at Club	1	61
April 30	Dinner at Club	10	71
May 1	Breakfast	1 6	72
"	Tea	6	78
14	Tea	1 1	79
June 1	Tea	1	80
			<u>80</u>

Jerry's super blurry camera

- One pixel, 1-bit color sensor (green=trees, white=snow)
- Model T: comes with 1-bit temperature sensor (hot, cold)

Basics: **Marginal** Probability

$$P(X = a) = \sum_b P(X = a, Y = b)$$

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$[P(\text{hot}), P(\text{cold})] = [\frac{195}{365}, \frac{170}{365}]$$

Probability Tables

- Write our distributions as tables
- # of entries? 4.
 - If we have n variables with k values, we get k^n entries
 - **Big!** For a 1080p screen, 12 bit color, size of table: $10^{7490589}$
 - No way of writing down all terms



Independence

- Independence between RVs:

$$P(X, Y) = P(X)P(Y)$$

- Example: simultaneously toss a coin and roll a die
- Why useful? Go from k^n entries in a table to $\sim kn$
- Expresses joint as **product** of marginals
- requires domain knowledge

Conditional Probability

- For when we know something (i.e. $Y=b$)

$$P(X = a|Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$

	green	white
hot	150/365	45/365
cold	50/365	120/365

$$P(cold|white) = \frac{P(cold,white)}{P(white)} = \frac{120}{45+120} = 0.73$$

Conditional independence

Same as independence, but conditioned on something

- require domain knowledge

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

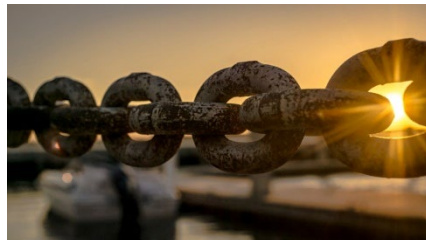
Chain Rule

- Apply repeatedly,

$$P(A_1, A_2, \dots, A_n)$$

$$= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1) \dots P(A_n|A_{n-1}, \dots, A_1)$$

- Note: still big!
 - If some **conditional independence**, can factor!
 - Leads to **probabilistic graphical models**



Break & Quiz

Q 2.1: Given joint distribution table:

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

What is the probability the temperature is hot given the weather is cloudy?

- A. $40/365$
- B. $2/5$
- C. $3/5$
- D. $195/365$

Break & Quiz

Q 2.1: Back to our joint distribution table:

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

What is the probability the temperature is hot given the weather is cloudy?

A. $40/365$

B. $2/5$

C. $3/5$

D. $195/365$

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

- A. 0.3
- B. 0.06
- C. 0.24
- D. 0.2

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

- A. 0.3
- B. 0.06
- C. 0.24
- D. 0.2**

Reasoning With Conditional Distributions

- Evaluating probabilities:
 - Wake up with a sore throat.
 - Do I have the flu?
- Logic approach: $S \rightarrow F$
 - Too strong.
- **Inference:** compute probability given evidence $P(F|S)$
 - Can be much more complex!



Bayes' Rule

Theorem: For any events A and B we have

$$P(A|B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Proof: Apply the chain rule two different ways:

$$\left. \begin{aligned} P(A, B) &= P(A | B) \cdot P(B) \\ &= P(B | A) \cdot P(A) \end{aligned} \right\} P(A|B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

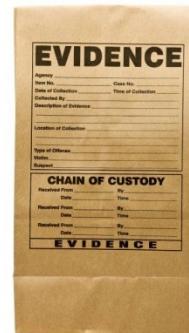
Using Bayes' Rule

- Want: $P(F|S)$
- **Bayes' Rule:** $P(F|S) = \frac{P(F,S)}{P(S)} = \frac{P(S|F)P(F)}{P(S)}$
- Parts:
 - $P(S) = 0.1$ Sore throat rate
 - $P(F) = 0.01$ Flu rate
 - $P(S|F) = 0.9$ Sore throat rate among flu sufferers
- **So:** $P(F|S) = 0.09$

Using Bayes' Rule

- Interpretation $P(F|S) = 0.09$
 - Much higher chance of flu than normal rate (0.01).
 - Very different from $P(S|F) = 0.9$
 - 90% of folks with flu have a sore throat
 - But, only 9% of folks with a sore throat have flu

- Idea: **update** probabilities from
evidence



Bayesian Inference

- Fancy name for what we just did. Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- H is the hypothesis
- E is the evidence



Bayesian Inference


- Terminology:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \longleftarrow \text{Prior}$$

- Prior: estimate of the probability **without** evidence

Bayesian Inference

- Terminology:



A black arrow points from the word "Likelihood" to the term $P(E|H)$ in the numerator of the equation.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Likelihood

- Likelihood: probability of evidence **given a hypothesis**

Bayesian Inference

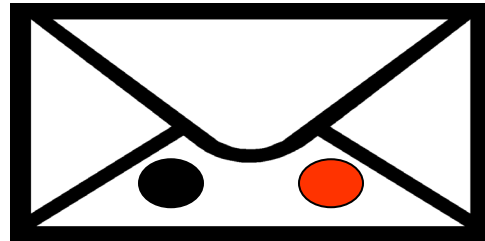
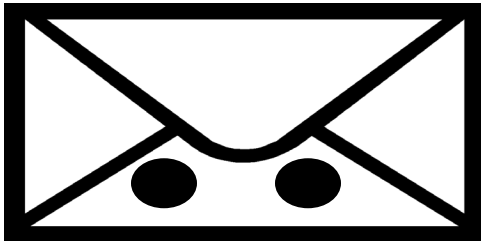
- Terminology:

$$\underset{\substack{\uparrow \\ \text{Posterior}}}{P(H|E)} = \frac{P(E|H)P(H)}{P(E)}$$

- Posterior: probability of hypothesis **given evidence**.

Two Envelopes Problem

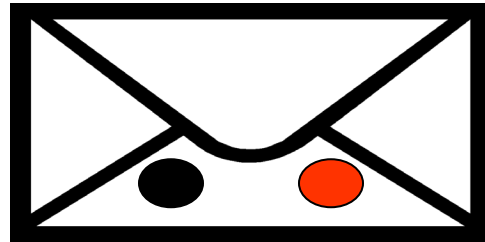
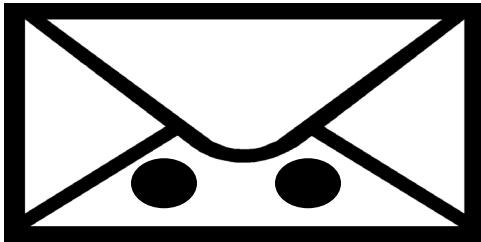
- We have two envelopes:
 - E_1 has two black balls, E_2 has one black, one red
 - The **red** one is worth \$100. Others, zero
 - Open an envelope, see one ball. Then, can switch (or not).
 - You see a black ball. **Switch?**



Two Envelopes Solution

- Let's solve it.
$$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$
- Now plug in:
$$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$
$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!



Naïve Bayes

- Conditional Probability & Bayes:

$$P(H|E_1, E_2, \dots, E_n) = \frac{P(E_1, \dots, E_n|H)P(H)}{P(E_1, E_2, \dots, E_n)}$$

- If we further make the **conditional independence assumption (a.k.a. Naïve Bayes)**

$$P(H|E_1, E_2, \dots, E_n) = \frac{P(E_1|H)P(E_2|H) \cdots P(E_n|H)P(H)}{P(E_1, E_2, \dots, E_n)}$$

Naïve Bayes

- Expression

$$P(H|E_1, E_2, \dots, E_n) = \frac{P(E_1|H)P(E_2|H) \cdots P(E_n|H)P(H)}{P(E_1, E_2, \dots, E_n)}$$

- H : some class we'd like to infer from evidence
 - We know prior $P(H)$
 - Estimate $P(E_i|H)$ from data! (“training”)
 - Very similar to envelopes problem.

Break & Quiz

Q 3.1: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

- A. $5/104$
- B. $95/100$
- C. $1/100$
- D. $1/2$

Break & Quiz

Q 3.1: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

- A. 5/104**
- B. 95/100
- C. 1/100
- D. 1/2

S : Spam

NS: Not Spam

DS: Detected as Spam

$P(S) = 50\%$ spam email

$P(NS) = 50\%$ not spam email

$P(DS|NS) = 5\%$ false positive, detected as spam but not spam

$P(DS|S) = 99\%$ detected as spam and it is spam

Applying Bayes Rule

$$P(NS|DS) = (P(DS|NS)*P(NS)) / P(DS) = (P(DS|NS)*P(NS)) / (P(DS|NS)*P(NS) + P(DS|S)*P(S)) = 5/104$$

Break & Quiz

Q 3.2: A fair coin is tossed three times. Find the probability of getting 2 heads and a tail

- A. $1/8$
- B. $2/8$
- C. $3/8$
- D. $5/8$

Break & Quiz

Q 3.2: A fair coin is tossed three times. Find the probability of getting 2 heads and a tail

A. $1/8$

B. $2/8$

C. $3/8$

D. $5/8$



Statistics

Bayesian Inference

- Conditional Probability & Bayes Rule:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- Evidence E : what we can observe
- Hypothesis H : what we'd like to infer from evidence
 - Need to plug in prior, likelihood, etc.
- Usually do not know these probabilities. How to estimate?

Samples and Estimation

- Usually, we don't know the distribution P
 - Instead, we see a bunch of samples
- Typical statistics problem: **estimate distribution** from samples
 - Estimate probabilities $P(H)$, $P(E)$, $P(E|H)$
 - Estimate the mean $E[X]$
 - Estimate parameters $P_{\theta}(X)$



Samples and Estimation

- Example: Bernoulli with parameter p
(*i.e. a weighted coin flip*)
 - $P(X = 1) = p$
 - Mean $E[X]$ is p



Examples: Sample Mean

- Bernoulli with parameter p
- See samples x_1, x_2, \dots, x_n
 - Estimate mean with **sample mean**

$$\hat{\mathbb{E}}[X] = \frac{1}{n} \sum_{i=1}^n x_i$$

- That is, counting heads



Break & Quiz

Q 2.1: You see samples of X given by $[0,1,1,2,2,0,1,2]$. Empirically estimate $\mathbb{E}[X^2]$

- A. $9/8$
- B. $15/8$
- C. 1.5
- D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by $[0,1,1,2,2,0,1,2]$. Empirically estimate $\mathbb{E}[X^2]$

A. $9/8$

B. $15/8$

C. 1.5

D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

Break & Quiz

Q 2.1: You see samples of X given by $[0,1,1,2,2,0,1,2]$. Empirically estimate $\mathbb{E}[X^2]$

A. $9/8$

B. $15/8$

C. 1.5

D. There aren't enough samples to estimate $\mathbb{E}[X^2]$

$$\begin{aligned} E[X^2] &\approx \frac{1}{n} \sum_i X_i^2 \\ &= \frac{1}{8} (0^2 + 1 + 1 + 4 + 4 + 0 + 1 + 4) = 15/8 \end{aligned}$$

Estimating Multinomial Parameters

- k -sized die (special case: $k=2$ coin)
- Face i has probability p_i , for $i=1\dots k$
- In n rolls, we observe face i showing up n_i times

$$\sum_{i=1}^k n_i = n$$

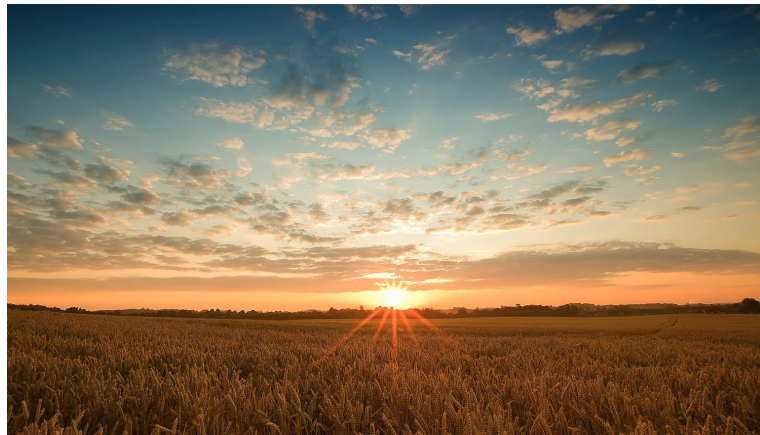
- Estimate (p_1, \dots, p_k) from this data (n_1, \dots, n_k)

Maximum Likelihood Estimate (MLE)

- The MLE of multinomial parameters $(\widehat{p}_1, \dots, \widehat{p}_k)$

$$\widehat{p}_i = \frac{n_i}{n}$$

- Estimate using frequencies



Break & Quiz

Q 2.2: You are empirically estimating $P(X)$ for some random variable X that takes on 100 values. You see 50 samples. How many of your $P(X=a)$ estimates might be 0?

- A. None.
- B. Between 5 and 50, exclusive.
- C. Between 50 and 100, inclusive.
- D. Between 50 and 99, inclusive.

Break & Quiz

Q 2.2: You are empirically estimating $P(X)$ for some random variable X that takes on 100 values. You see 50 samples. How many of your $P(X=a)$ estimates might be 0?

For each a , your estimate is $P(X = a) = \frac{\text{\#samples taking value } a}{50}$

- A. None.
- B. Between 5 and 50, exclusive.
- C. Between 50 and 100, inclusive.
- D. Between 50 and 99, inclusive.**

If you don't see a number at all in the 50 samples then the estimated probability of that number is 0.

You can see up to 50 different values in 50 samples. On the other hand, all 50 samples might have the same value in which case 99 values were never seen.

Regularized Estimate

- Hyperparameter $\epsilon > 0$

$$\hat{p}_i = \frac{n_i + \epsilon}{n + k\epsilon}$$

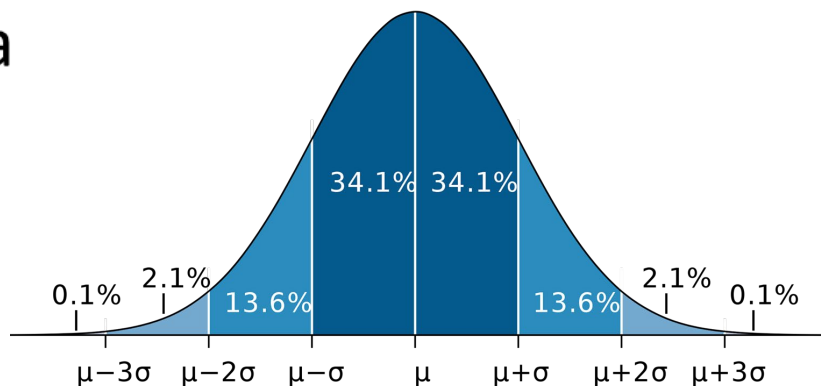
- Avoids zero when n is small
- Biased, but has smaller variance
- Equivalent to a specific Maximum A Posteriori (MAP) estimate, or smoothing

Estimating 1D Gaussian Parameters

- Gaussian (aka Normal) distribution $N(\mu, \sigma^2)$
 - True mean μ , true variance σ^2
- Observe n data points from this distribution

$$x_1, \dots, x_n$$

- Estimate μ, σ^2 from this data



Estimating 1D Gaussian Parameters

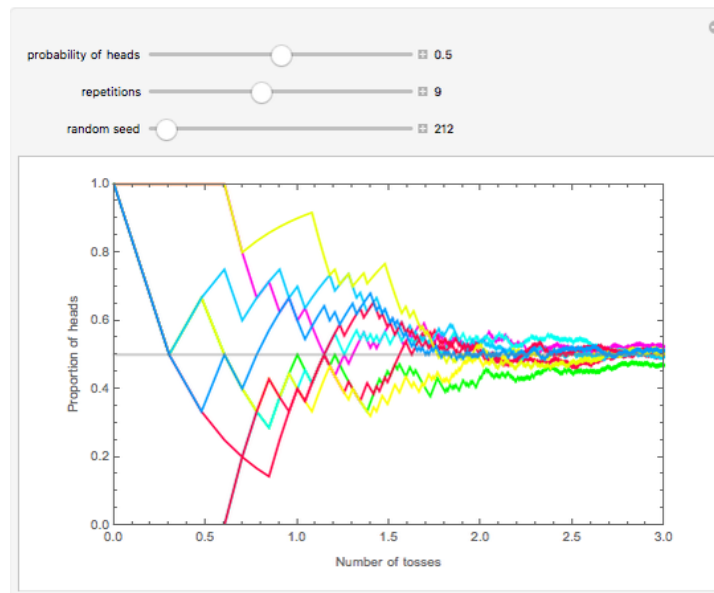
- Mean estimate $\hat{\mu} = \frac{x_1 + \dots + x_n}{n}$
- Variance estimates

- Unbiased $s^2 = \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n - 1}$

- MLE $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n}$

Estimation Theory

- Is the sample mean a good estimate of the true mean?
 - Law of large numbers
 - Central limit theorems



Wolfram Demo

Estimation Errors

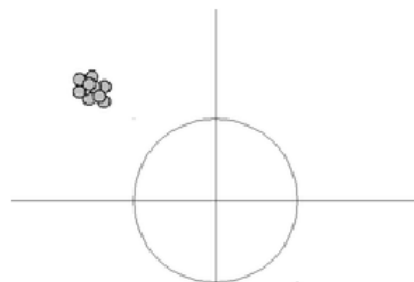
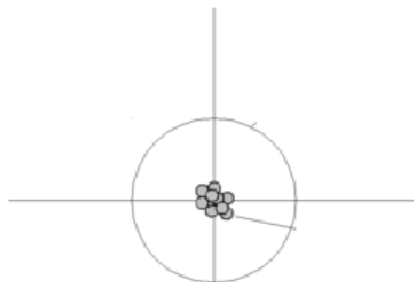
- With finite samples, likely error in the estimate.
- Mean squared error
 - $\text{MSE}[\hat{\theta}] = \mathbb{E} [(\hat{\theta} - \theta)^2]$
- Bias / Variance Decomposition
 - $\text{MSE}[\hat{\theta}] = \underbrace{\mathbb{E} [(\hat{\theta} - E[\hat{\theta}])^2]}_{\text{Variance}} + \underbrace{(E[\hat{\theta}] - \theta)^2}_{\text{Bias}}$

Bias / Variance

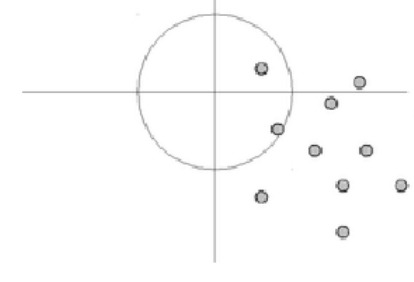
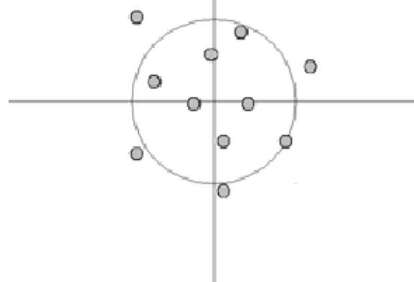
Low Bias

High Bias

Low Variance



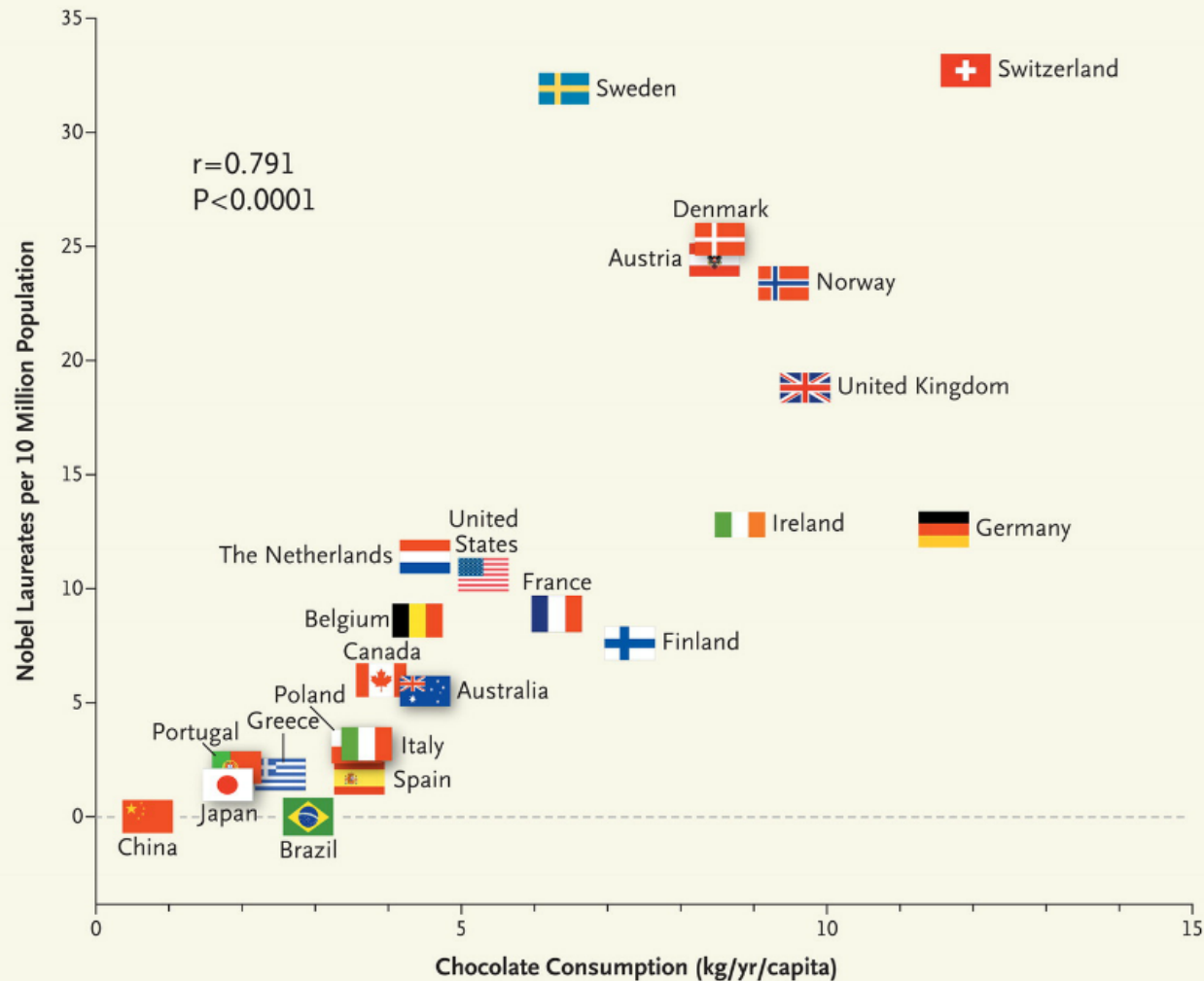
High Variance



Wikipedia: Bias-variance tradeoff

Correlation vs. Causation

- Conditional probabilities only define correlation (aka association)
- $P(Y|X)$ “large” does not mean X causes Y
- Example: X =yellow finger, Y =lung cancer
- Common cause: smoking



Readings

- Vast literature on intro probability and statistics.
- Local classes: **Math/Stat 431**
- **Suggested reading:**
Probability and Statistics: The Science of Uncertainty,
Michael J. Evans and Jeff S. Rosenthal
<http://www.utstat.toronto.edu/mikevans/jeffrosenthal/book.pdf>

(Chapters 1-3, excluding “advanced” sections)