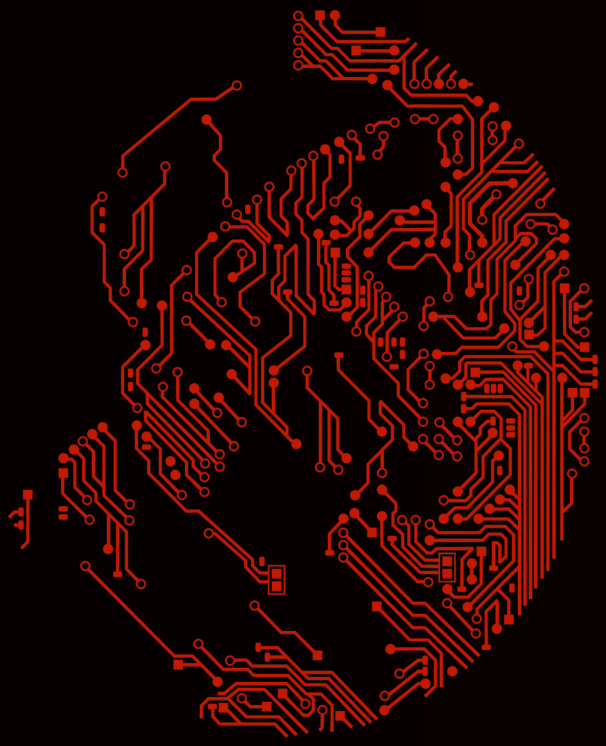
# MAC0459/MAC5865



# Data Science and Engineering

https://app.sli.do/event/c0i1mpf6

18th class

Roberto Hirata Jr
hirata@ime.usp.br

# Information Theory

# Informação

- A aula já começou !

# Informação

- H luaa ej ça kbou !

# Informação

- O quê diz a primeira frase ?
  - A aula já começou !
- E a segunda ?
  - H luaa ej ça kbou !
- Qual das duas contém mais informação ?

# Conceito de Informação

- Infomação é definida em relação a uma mensagem.
- Uma mensagem pode ser qualquer coisa que é enviada por um emissor para um receptor específico
  - Exemplo, uma imagem
- Quando abstraimos os aspectos físicos de uma mensagem, sobra a informação.

# Conceito de Informação

- O emissor e o receptor podem ser considerados como entidades abstratas
  - Suponha que o emissor é capaz de gerar uma mensagem (uma imagem, um sinal)
  - Suponha que o receptor é capaz de ler a mensagem (por exemplo, carregar na memória)

# Conceito de Informação

- A informação é o “conteúdo” da mensagem
- Não o suporte físico da mesma
- Por exemplo, na frase:
  - A aula já começou !
    - **Seu suporte físico são as letras, as palavras, escritas e apresentadas**
    - **Seu conteúdo é aquilo que elas significam, a ideia de aula, o conceito de começar, o fato de ela já ter ou não começado e a exclamação no final.**
    - **A questão de se a aula já havia começado ou não quando a mensagem foi recebida, muda o conteúdo de informação da mensagem**

# Conceito de Informação

- Portanto, o estado do receptor é significativo para se determinar as características da informação
  - Se a aula ainda não tivesse começado no entender dos estudantes, a mensagem poderia ter o significado: “Vou começar agora”.
  - Se os estudantes entendessem que a aula já havia começado, poderiam interpretar a mensagem como: “Vocês não estão prestando atenção”.

# Conceito de Informação

- A informação permite alterar o estado de conhecimento do receptor sobre o emissor

emissor


receptor

**Não faz a menor ideia**

**Pode imaginar o que seja**

**Imagina aquilo que gostaria**

**Imaginou aquilo que nem deveria**

# Conceito de Informação

- Quanto mais improvável a mensagem, mais informação ela transporta

emissor

*...não sabe o que eu tenho pra te contar ...*

*O que será?!!*

receptor

**Não faz a menor ideia**

**Pode imaginar o que seja**

**Imagina aquilo que gostaria**

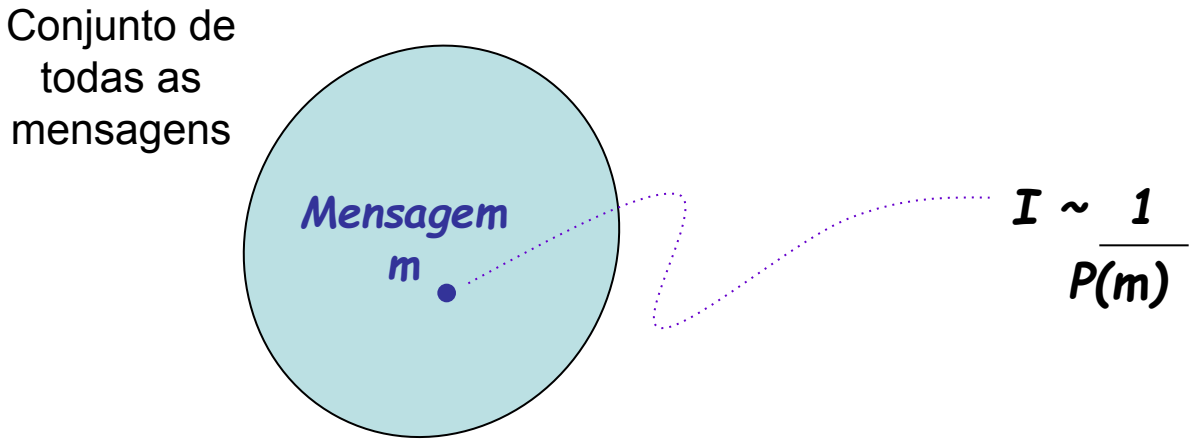
**Imaginou aquilo que nem deveria**

# Medida da Informação

- Resumindo:
  - Quanto mais provável a mensagem (i.e., seu conteúdo), tanto menor a informação que ela traz
- Com base nesse conceito vamos “construir” uma medida para a informação
- Em seguida, vamos interpretar essa medida assim construída

# Medida da informação

- A quantidade de informação em uma mensagem depende inversamente da probabilidade da mensagem
  - Em nosso caso a mensagem é a imagem



# Medida da informação

- Quando se tem a conjunção de duas mensagens :
  - Do ponto de vista lógico booleano, é uma operação “e” (AND)
  - Do ponto de vista de teoria de conjuntos é uma intersecção
- Exemplo: se você envia a mesma mensagem duas vezes, deve ser entendida como uma só:

$$M.M = M$$

# Medida da informação

- Quando se tem a conjunção de duas mensagens :
  - Desejamos que a informação seja **aditiva** nesse caso, i.e., dadas duas mensagens concomitantes, queremos que suas informações se somem
- Porém, nesse caso suas probabilidades se **multiplicam**:

$$P(m1 \bullet m2) = P(m1).P(m2)$$

Conjunção (AND)

Multiplicação

# Medida da informação

- Estabelecemos que a medida da informação deve somar quando as probabilidades se multiplicam
  - A solução funcional para isso é empregar o logaritmo:

$$I \sim \frac{1}{P(m)}, \quad \left.\begin{array}{l} I(m1 \cdot m2) = I(m1) + I(m2) \\ P(m1 \cdot m2) = P(m1).P(m2) \end{array}\right\} \quad I = \log\left(\frac{1}{P(m)}\right)$$

Conjunção (AND)

Multiplicação

# Medida da informação

$$I \sim \frac{1}{P(m)} \qquad I(m1 \cdot m2) = I(m1) + I(m2) \qquad I = \log \left( \frac{1}{P(m)} \right)$$

ou $I = - \log P(m)$

portanto $I(m1 . m2) = - \log P(m1 . m2) = -\log [ P(m1) . P(m2) ]$

$= -\log P(m1) - \log P(m2) = I(m1) + I(m2)$

# Medida da Informação

- Interpretemos agora a medida da informação que acabamos de construir seguindo um procedimento heurístico

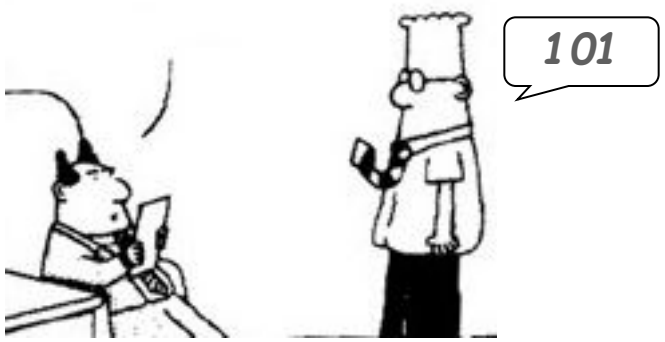
$$I = - \log P(m)$$

- Primeiramente, vamos nos ater a um caso limite, em que as mensagens são simples e equiprováveis

# Medida da Informação

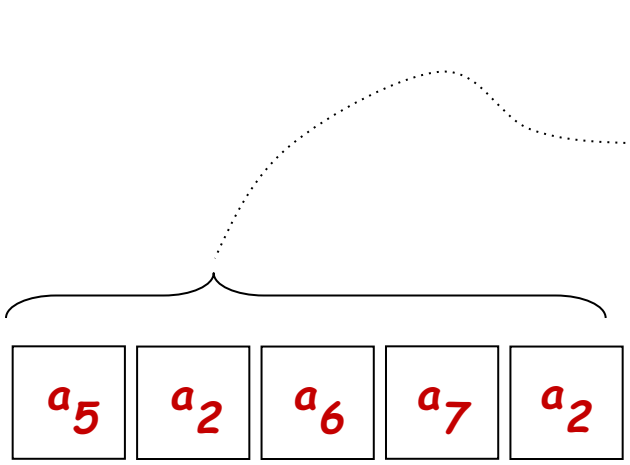
- Suponhamos que todas as mensagens sejam equiprováveis. Vamos ordená-las em ordem lexicográfica e atribuir a cada uma delas um número.

COMBINEMOS OS CÓDIGOS DAS SMS'S :
MENSAGEM 1 – AR-CONDICIONADO QUEBRADO
MENSAGEM 2 – BANCO DE DADOS FORA DO AR
MENSAGEM 3 – COMPUTADORES SEM REDE
MENSAGEM 4 – ... NÃO ENTENDI ...
ENTENDEU ?




**Se tivermos ao todo 10 diferentes mensagens e usarmos 10 números para representá-las, quantos dígitos serão necessários ?**

# Medida da Informação




Se cada número de **d** posições preenchidas com algarismos do conjunto {0,1,...,**a**-1} simbolizar uma mensagem, ao todo quantas **M** mensagens teremos ?

$$M = a^d$$

# Medida da Informação

- O resultado significa que com **a** símbolos e **d** dígitos pode-se representar $\mathbf{M = a^d}$ mensagens.
- Inversamente: sabendo-se que temos **M** mensagens e **a** símbolos, quantos dígitos **d** são necessários para representar essas mensagens ?
  - Resposta: $\mathbf{d = \log_a M}$

# Medida da Informação

- Exemplo:
  - Se **a = 2** algarismos temos: símbolos = {0,1}
    □ conjunto binário (apenas 2 símbolos)
  - Quantos dígitos binários (i.e., bits) serão necessários para representar **M** mensagens ?

$$\mathbf{d = \log_2 M}$$

- Portanto se usarmos log na base 2, a resposta será dada em bits.

# Medida da Informação

- Suponhamos que as M mensagens sejam equiprováveis. Então, qual a probabilidade de cada mensagem $m_i$ ?

$$P(m_i) = \frac{1}{M}$$

- Logo:

$$M = \frac{1}{P(m_i)}$$

# Medida da Informação

Portanto, uma mensagem $\mathbf{m_i}$ de um conjunto de **M** mensagens equiprováveis pode ser representada com **a** símbolos requerendo **d** dígitos:

$$M = \frac{1}{P(m_i)}, \quad d = \log_a M \quad \square \quad d = \log_a \frac{1}{P(m_i)}$$

Então, a medida da informação da mensagem $\mathbf{m_i}$ de probabilidade $\mathbf{P(m_i)}$ pode ser interpretada como o comprimento em dígitos requerido para representar a mensagem com **a** símbolos

$$I_i = - \log_a P(m_i)$$

# Entropia

- Define-se a **entropia** de um conjunto de símbolos (mensagem) como a **informação média** dos símbolos presentes na mensagem.
- Por ser informação média, a entropia mede a quantidade média de dígitos por símbolo, requeridos para representar a mensagem.

# Entropia

- Para cada símbolo $\mathbf{m_i}$ :

$$\mathbf{I_i = - \log_a P(m_i)}$$

- Logo, para uma mensagem com **n** símbolos:

$$\mathbf{H = \langle I \rangle = - \sum_{i=0}^{n-1} \log_a P(m_i) \, . \, P(m_i)}$$

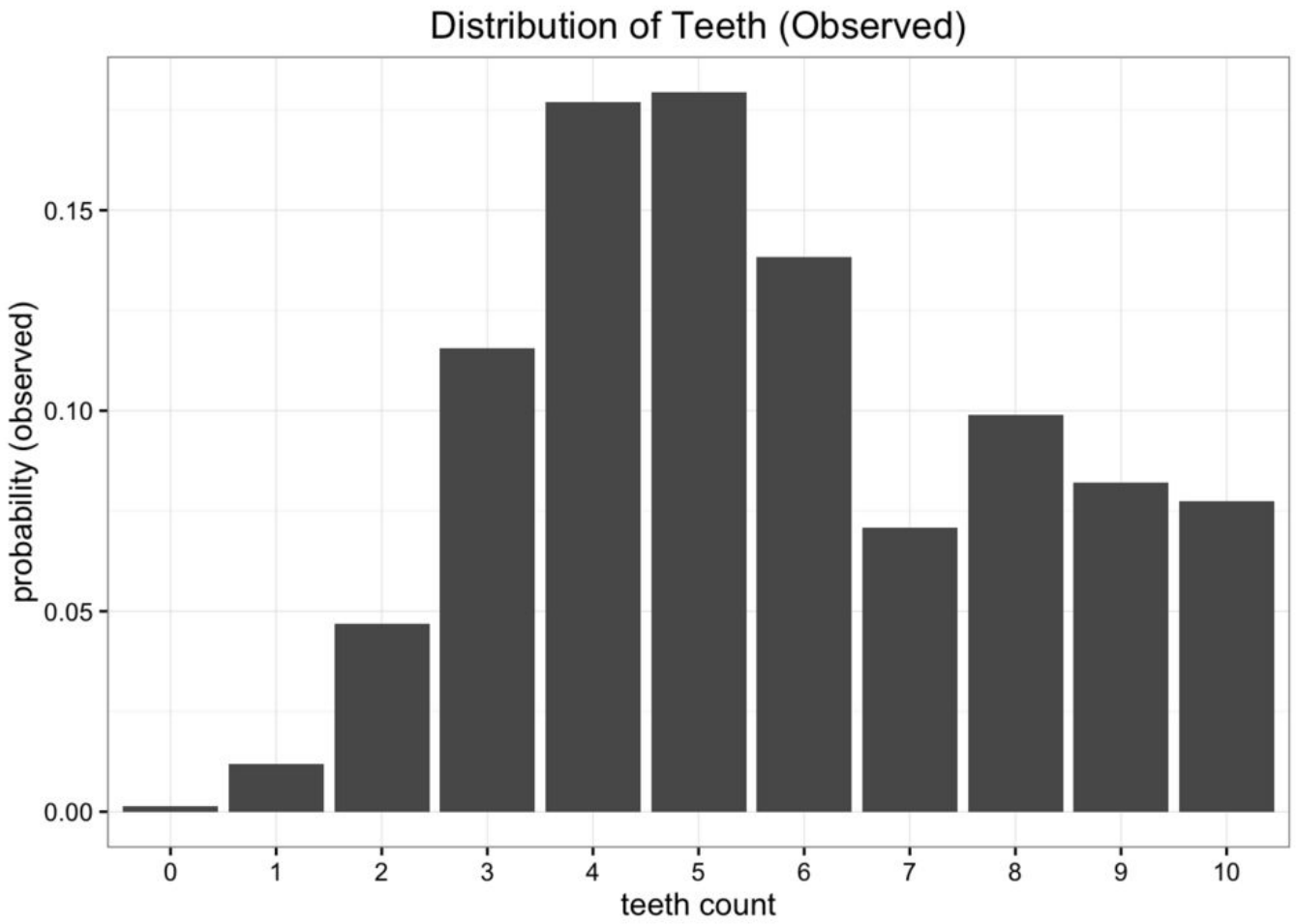
# Entropy

- Given a discrete random variable $X = \{x_1, x_2, ..., x_n\}$, the entropy of X is given by:

$$H = \langle X \rangle = -\sum_{i=1}^{n} \log_a P(x_i) \cdot P(x_i)$$

- If $P(x_i) = 0$, $I_i = 0$ because $\lim_{x \to 0} \log_a P(x_i) \cdot P(x_i) = 0$

# Entropy computation




Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Entropy

- Entropy estimation is a real problem
- Usually we underestimate if we use the naive formula
- There are methods to better estimate the entropy

# Entropy

- Computation of the entropy of the gray-levels of an image (notebook)

# Conditional Entropy

- Given two discrete random variables $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_m\}$ the conditional entropy of Y given X is:

$$H(Y|X) = -\sum_{i,j=1}^{n,m} \log_a P(y_j|x_i) \cdot P(x_i, y_j)$$

# Mutual Information

- Given two discrete random variables X and Y the mutual information of X and Y is:

$$\mathbf{I}(X;Y) = \mathbf{H}(X) - \mathbf{H}(Y|X) = \mathbf{H}(Y) - \mathbf{H}(X|Y)$$

# Kullback-Leibler Divergence

- Compare two distributions ***P*** and ***Q***:

$$D_{KL}(P||Q) = \sum_{i=0}^{n-1} \left( \log_a P(m_i) - \log_a Q(m_i) \right) . P(m_i)$$

$$D_{KL}(P||Q) = \sum_{i=0}^{n-1} P(m_i) \left( \log_a P(m_i)/Q(m_i) \right)$$

# Kullback-Leibler Divergence

- Properties

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

$$I(X;Y) = D_{KL}(P(X,Y)||P(X)P(Y))$$

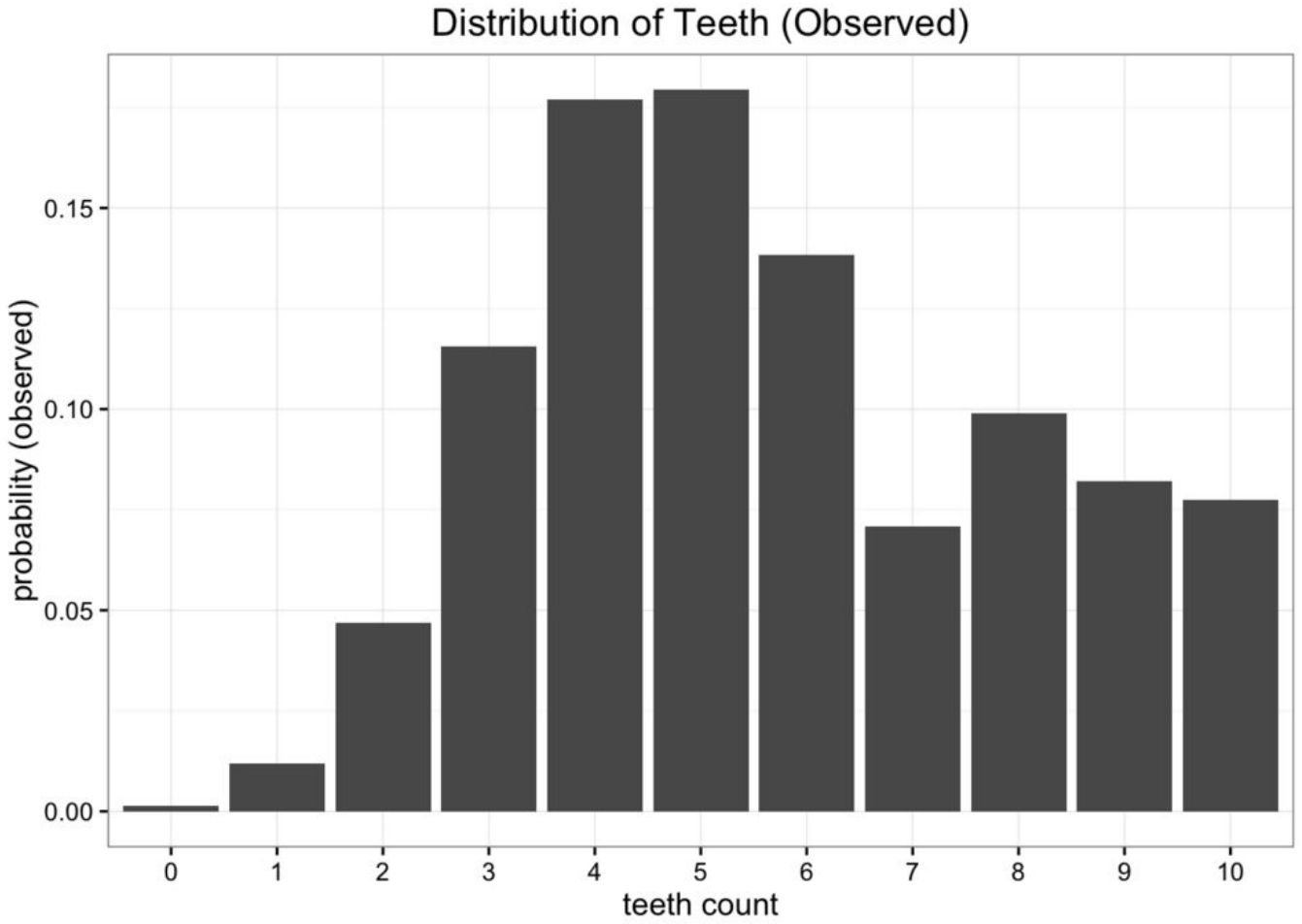
- Cross entropy

$$H(P,Q) = H(P) + D_{KL}(P||Q)$$

$$H(P,Q) = -\sum \log_a Q(x) \cdot P(x)$$

# Kullback-Leibler Divergence

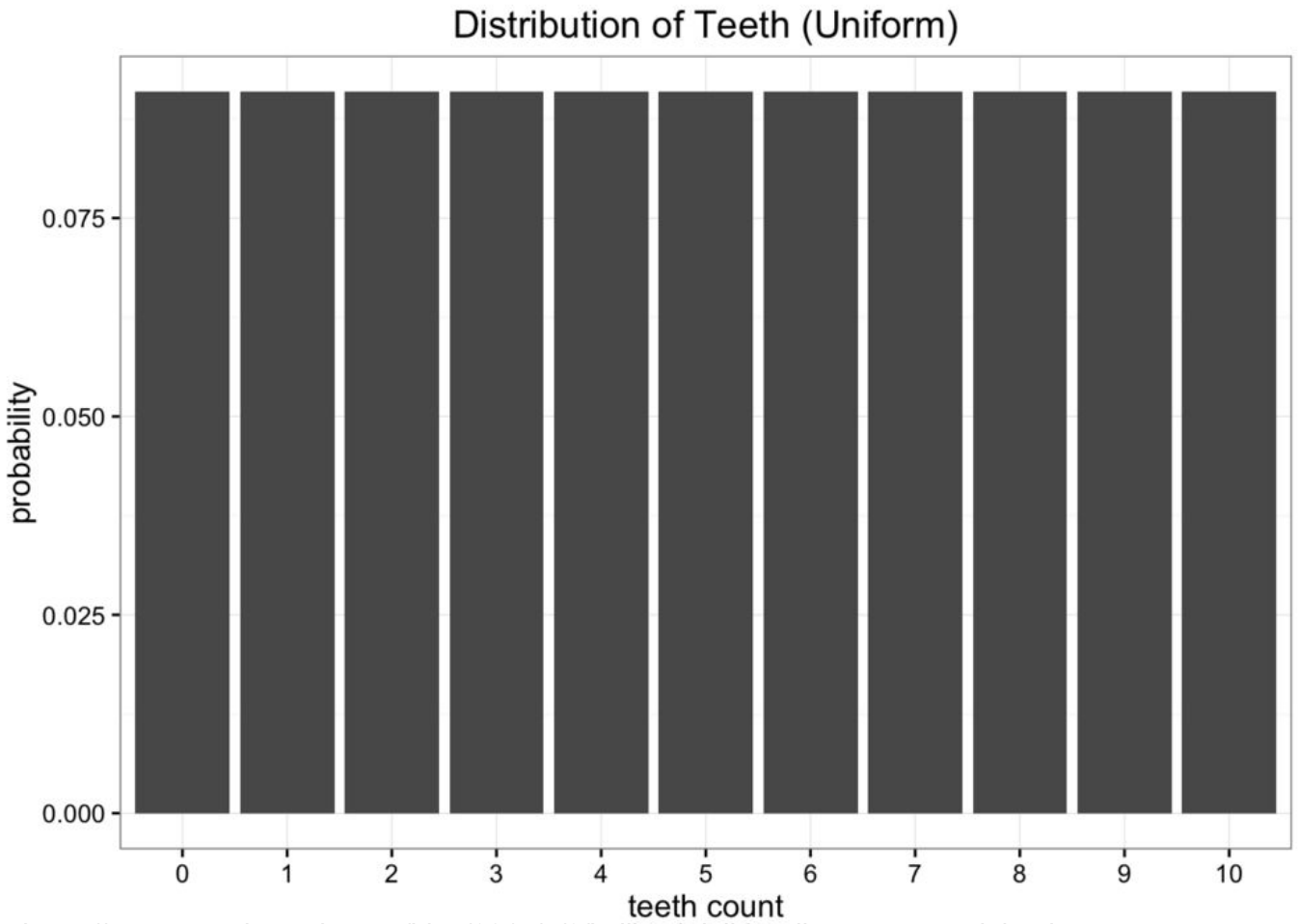
- Compare two distributions ***P*** and ***Q***:




Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Kullback-Leibler Divergence

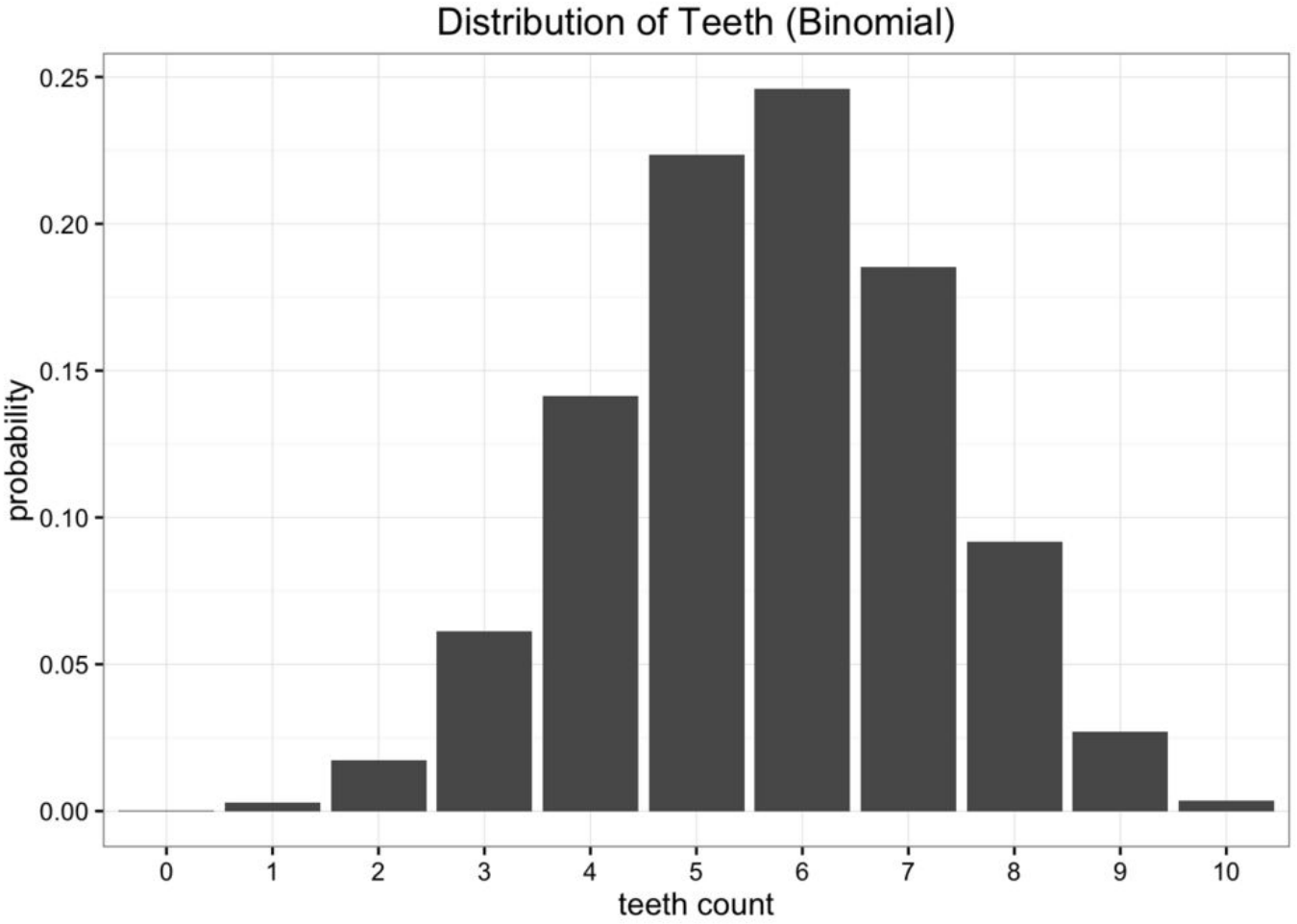
- Compare two distributions ***P*** and ***Q***:




Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Kullback-Leibler Divergence

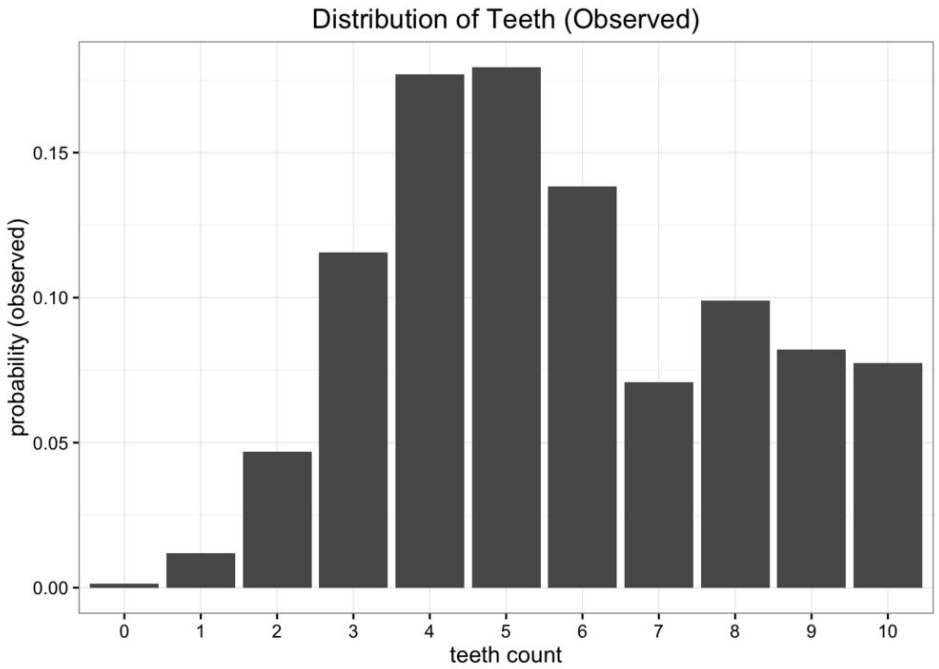
- Compare two distributions ***P*** and ***Q***:

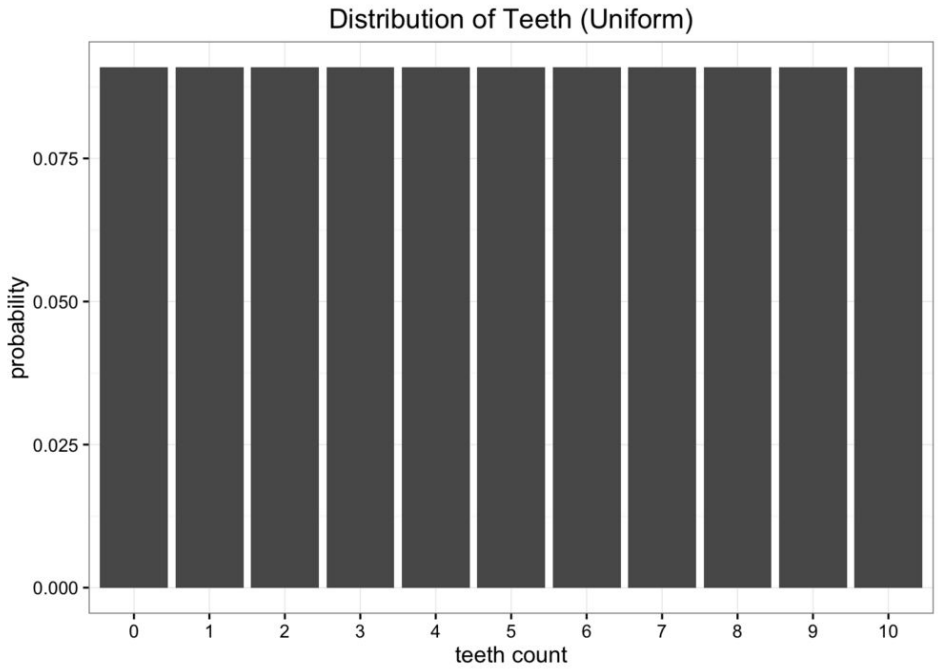



Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Kullback-Leibler Divergence

- Compare two distributions ***P*** and ***Q***:

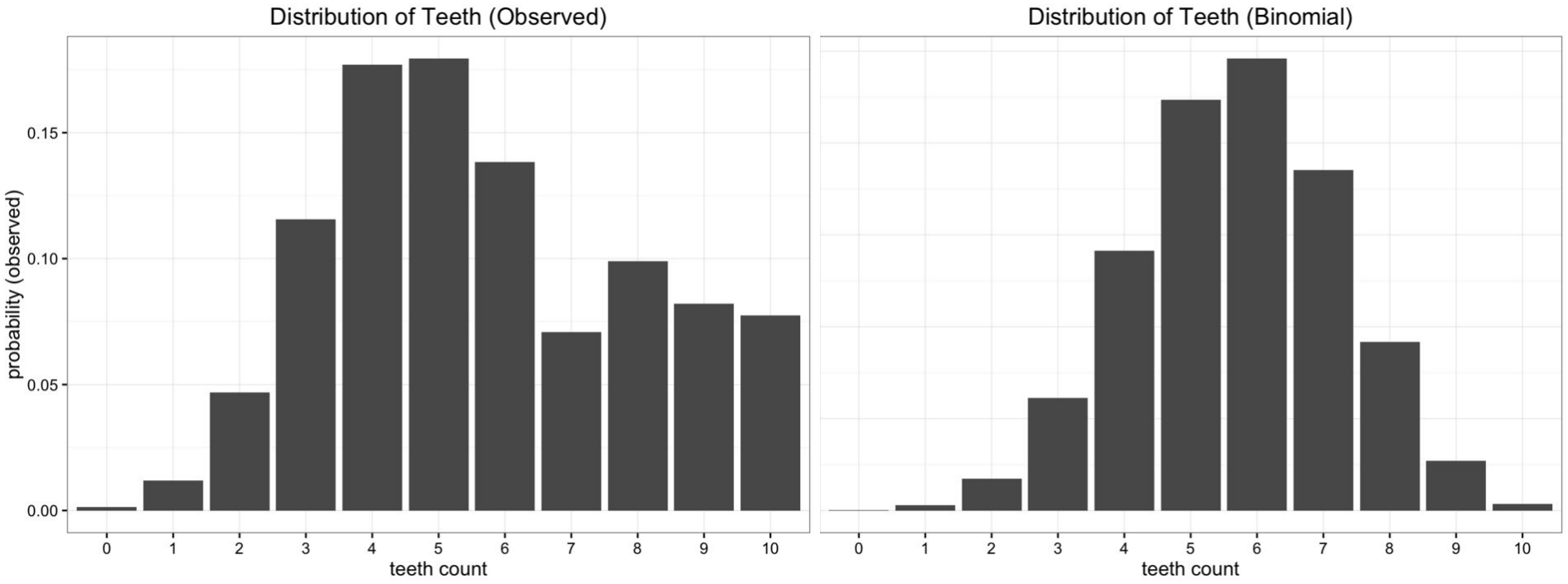



Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Kullback-Leibler Divergence

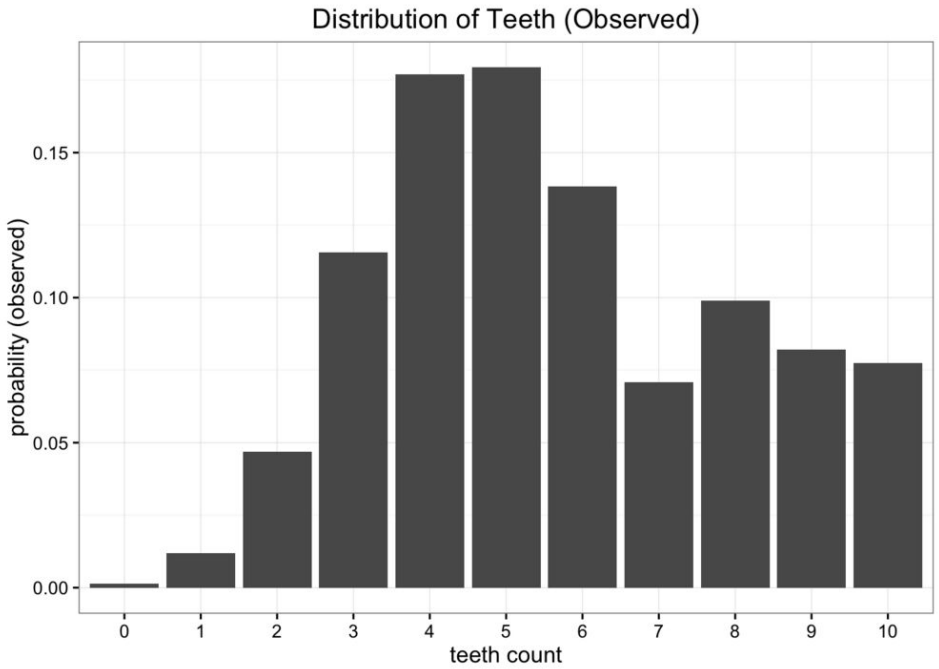
- Compare two distributions ***P*** and ***Q***:

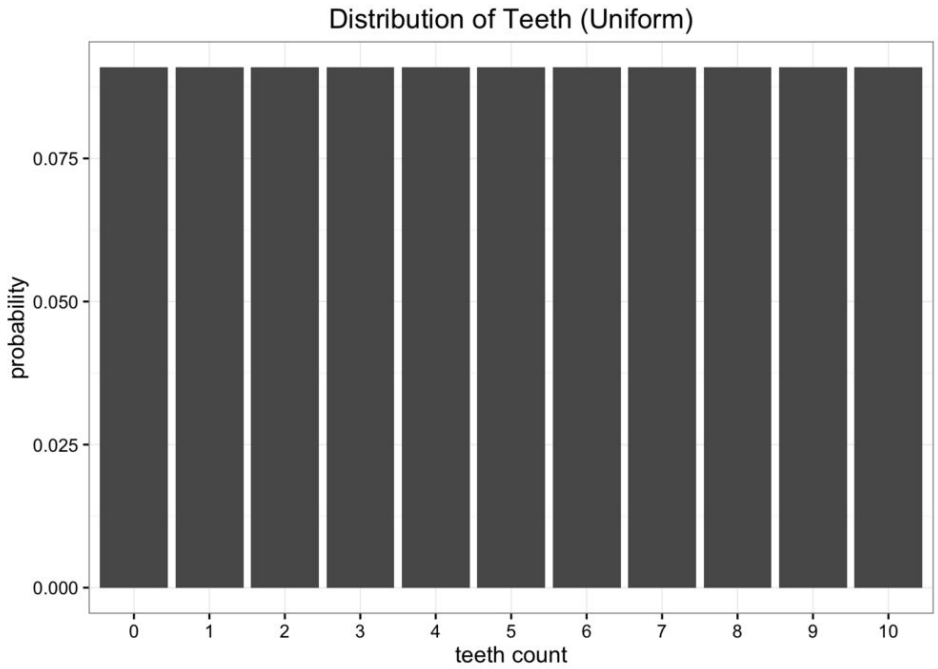



Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Kullback-Leibler Divergence

- Compare two distributions ***P*** and ***Q***:



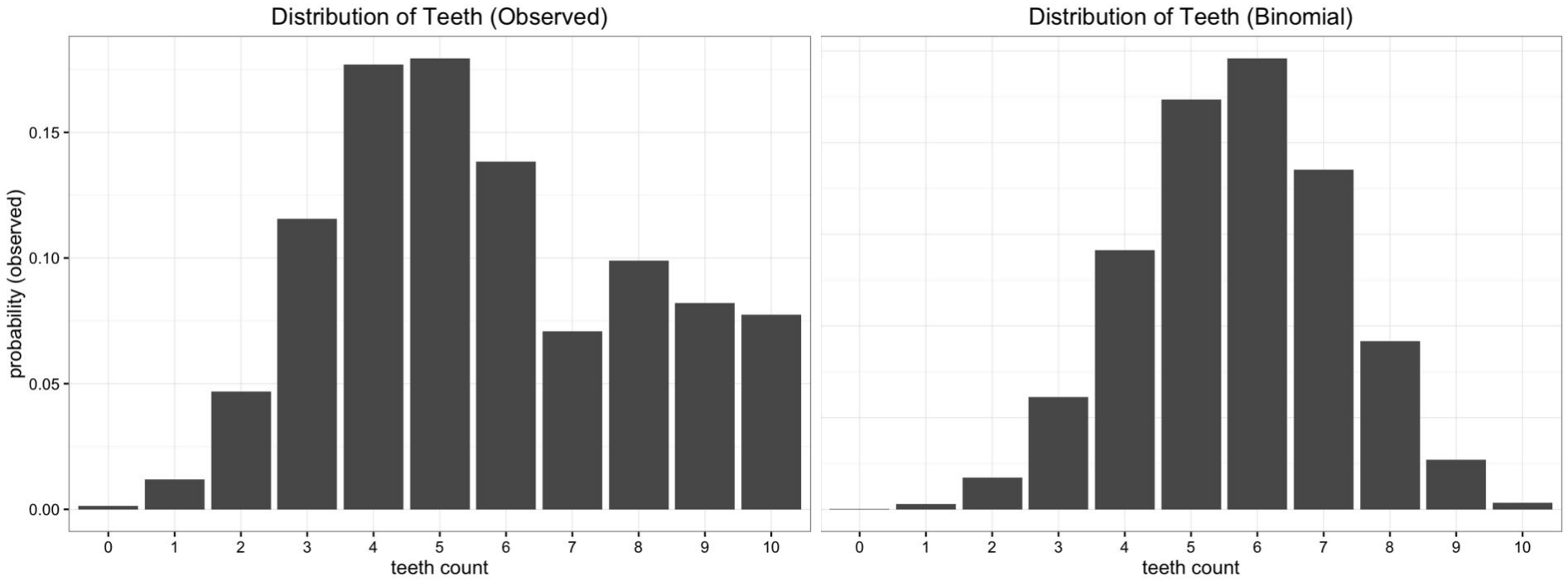




$D_{KL}(Observed||Uniform) = 0.338$

Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained

# Kullback-Leibler Divergence

- Compare two distributions ***P*** and ***Q***:




$D_{KL}(Observed||Binomial) = 0.477$

Source:https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained