The Pokemon Collector's Problem: What are the Chansey?

Kevin Wang

School of Mathematics and Statistics
The University of Sydney

October 28, 2016

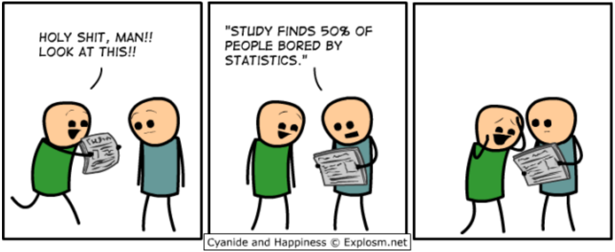
Motivation

- Interview question
- Statistical tools
  - Probability (theoretical)
  - Simulation (methodological)
  - Data analysis (applied)

- To keep myself entertained and keep my PhD supervisors "amused".

Motivation

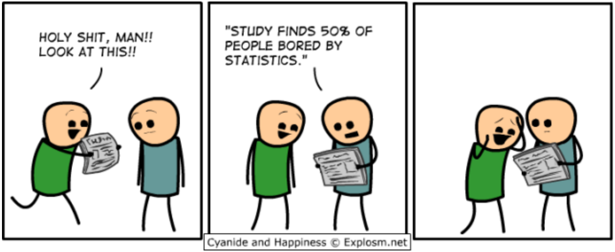
- Interview question
- Statistical tools
  - Probability (theoretical)
  - Simulation (methodological)
  - Data analysis (applied)



- To keep myself entertained and keep my PhD supervisors "amused".

Motivation

- Interview question
- Statistical tools
  - Probability (theoretical)
  - Simulation (methodological)
  - Data analysis (applied)



- To keep myself entertained and keep my PhD supervisors "amused".

The Pokemon Collector's Problem

Problem

There are 142 individual Pokemons on the game "Pokemon GO". Assuming we can collect 1 Pokemon per unit time, what is the expected number of Pokemons do we need to complete the collection?

- According to report [1], one player caught 4,269 Pokemons to complete this collection.
- I am on 1,430 in total, 99 of them are unique. And all of my PhD supervisors are amused/annoyed/urging me to go back to research.

[1] http://www.polygon.com/2016/7/21/12247678/pokemon-go-complete-pokedex

The Pokemon Collector's Problem

Problem

There are 142 individual Pokemons on the game "Pokemon GO". Assuming we can collect 1 Pokemon per unit time, what is the expected number of Pokemons do we need to complete the collection?

- According to report [1], one player caught 4,269 Pokemons to complete this collection.
- I am on 1,430 in total, 99 of them are unique. And all of my PhD supervisors are amused/annoyed/urging me to go back to research.

[1] http://www.polygon.com/2016/7/21/12247678/pokemon-go-complete-pokedex

The Problem is hard

- Pokemon problem has:
  - unequal probabilities
  - multiple collection with differing termination point (evolution)
  - arrival in batches of different sizes (eggs)

A theoretical closed form does not exist in literature so far.

The Coupon Collector's Problem (CCP)

- It is a weaker version of the Pokemon problem.
- Classical elements of the problem include:
  - Equal/unequal probabilities
  - Single/multiple collection
  - Arrival in batches
- Some references:
  - *The Coupon Collector's Problem*, M. Ferrante, M. Saltalamacchia, (2014)
  - *Introduction to Probability Models.* S.Ross.
  - *A First Course in Probability.* S. Ross.

Key of this talk

1. A diverse range of possible solutions, combining a whole spectrum of Statistics.
2. Characteristics of the problems, including probability sensitivity, capture/evolution tradeoff.

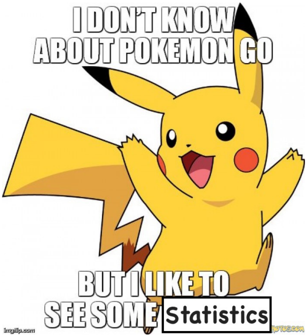


Figure: A Pikture without context.

Introduction

Single Collection
- Equal Probability Assumption
  - Basic Probability Solution
  - Markov Chain Solution (Omitted)
- Unequal Probability Assumption
  - Basic Probability Solution
  - Sensitivity of Overall Collection
  - Poisson Process Solution (Omitted)

Other Generalisations

Pokemon Problem Simulations

Equal probability CCP

Problem

Suppose we buy one coupon per unit time. Assuming each coupon is equally likely to appear, how many coupons we have to purchase (on average) to complete a collection of n different coupons?

- Let X be the (random) number of coupons we need to collect to complete collection, which has n distinct types of coupons.
- Let p_i denote the probability of collecting coupon of type i, which equals to $1/n$ according to our equal probability assumption.
- Let X_i denote the **additional** number of coupons we need to collect, to pass from $i-1$ to i distinct types of coupons in our collection. $i = 1, 2, \ldots, n$.
- Hence, $X = X_1 + X_2 + \cdots + X_n$.

Equal probability CCP

Problem

Suppose we buy one coupon per unit time. Assuming each coupon is equally likely to appear, how many coupons we have to purchase (on average) to complete a collection of n different coupons?

- Let X be the (random) number of coupons we need to collect to complete collection, which has n distinct types of coupons.
- Let p_i denote the probability of collecting coupon of type i, which equals to $1/n$ according to our equal probability assumption.
- Let X_i denote the **additional** number of coupons we need to collect, to pass from $i-1$ to i distinct types of coupons in our collection. $i = 1, 2, \ldots, n$.
- Hence, $X = X_1 + X_2 + \cdots + X_n$.

Equal probability CCP

Problem

Suppose we buy one coupon per unit time. Assuming each coupon is equally likely to appear, how many coupons we have to purchase (on average) to complete a collection of n different coupons?

- Let X be the (random) number of coupons we need to collect to complete collection, which has n distinct types of coupons.
- Let p_i denote the probability of collecting coupon of type i, which equals to $1/n$ according to our equal probability assumption.
- Let X_i denote the **additional** number of coupons we need to collect, to pass from $i-1$ to i distinct types of coupons in our collection. $i = 1, 2, \ldots, n$.
- Hence, $X = X_1 + X_2 + \cdots + X_n$.

Equal probability CCP

Problem

Suppose we buy one coupon per unit time. Assuming each coupon is equally likely to appear, how many coupons we have to purchase (on average) to complete a collection of n different coupons?

- Let X be the (random) number of coupons we need to collect to complete collection, which has n distinct types of coupons.
- Let p_i denote the probability of collecting coupon of type i, which equals to $1/n$ according to our equal probability assumption.
- Let X_i denote the **additional** number of coupons we need to collect, to pass from $i-1$ to i distinct types of coupons in our collection. $i = 1, 2, \ldots, n$.
- Hence, $X = X_1 + X_2 + \cdots + X_n$.

Equal probability CCP

Problem

Suppose we buy one coupon per unit time. Assuming each coupon is equally likely to appear, how many coupons we have to purchase (on average) to complete a collection of n different coupons?

- Let X be the (random) number of coupons we need to collect to complete collection, which has n distinct types of coupons.
- Let p_i denote the probability of collecting coupon of type i, which equals to $1/n$ according to our equal probability assumption.
- Let X_i denote the **additional** number of coupons we need to collect, to pass from $i-1$ to i distinct types of coupons in our collection. $i = 1, 2, \ldots, n$.
- Hence, $X = X_1 + X_2 + \cdots + X_n$.

Solution for equal probability case

Problem (Reduced)

When $i-1$ distinct types of coupons have already been collected, what is the probability that a **new** *coupon will be of a distinct type be?*

- $X_1 = 1$.
- Every time, since $i-1$ of those were already chosen, we are left with only $n-(i-1)$ possible types to choose from.
- Probability to escape from this state is $\xi_i = \frac{n-(i-1)}{n}$

Each X_i is a geometric distributed random variable with probability parameter ξ_i.

Solution for equal probability case

Problem (Reduced)

When $i-1$ distinct types of coupons have already been collected, what is the probability that a **new** *coupon will be of a distinct type be?*

- $X_1 = 1$.
- Every time, since $i-1$ of those were already chosen, we are left with only $n-(i-1)$ possible types to choose from.
- Probability to escape from this state is $\xi_i = \frac{n-(i-1)}{n}$

Each X_i is a geometric distributed random variable with probability parameter ξ_i.

Solution for equal probability case

Problem (Reduced)

When $i-1$ distinct types of coupons have already been collected, what is the probability that a **new** *coupon will be of a distinct type be?*

- $X_1 = 1$.
- Every time, since $i-1$ of those were already chosen, we are left with only $n-(i-1)$ possible types to choose from.
- Probability to escape from this state is $\xi_i = \frac{n-(i-1)}{n}$

Each X_i is a geometric distributed random variable with probability parameter ξ_i.

Solution for equal probability case

Problem (Reduced)

When $i-1$ distinct types of coupons have already been collected, what is the probability that a **new** *coupon will be of a distinct type be?*

- $X_1 = 1$.
- Every time, since $i-1$ of those were already chosen, we are left with only $n-(i-1)$ possible types to choose from.
- Probability to escape from this state is $\xi_i = \frac{n-(i-1)}{n}$

Each X_i is a geometric distributed random variable with probability parameter ξ_i.

Solution for equal probability case

- First year statistics course: the expectation of a $Geo(\xi)$ random variable is $1/\xi$. Hence:

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) \\
&= \frac{1}{\xi_1} + \frac{1}{\xi_2} + \cdots + \frac{1}{\xi_n} \\
&= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \\
&= n \cdot H_n,
\end{aligned}$$

where H_n is the n-th Harmonic number.

Solution for equal probability case

- First year statistics course: the expectation of a $Geo(\xi)$ random variable is $1/\xi$. Hence:

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) \\ &= \frac{1}{\xi_1} + \frac{1}{\xi_2} + \cdots + \frac{1}{\xi_n} \\ &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \\ &= n \cdot H_n,\end{aligned}$$

where H_n is the n-th Harmonic number.

Some characteristics

- The waiting time for the very last coupon is always the longest.
- $2H_2 - 1H_1 = 2$, $100H_{100} - 99H_{99} = 6.177$.
- $142H_{142} \approx 786.19$.
- $\mathbb{E}(X) = nH_n = n\log n + \gamma n + \frac{1}{2} + o(1)$.
- $\mathbb{E}(X)$ grows at the rate of $O(n\log(n))$, which is not too bad.
- $Var(X) < \frac{\pi^2}{6}n^2$, Chebyshev inequality gives:

$$\mathbb{P}(|X - nH_n| \geq cn) \leq \frac{\pi^2}{6c^2}. \tag{1}$$

Some characteristics

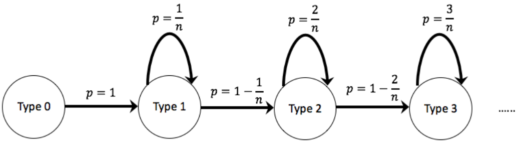
- The waiting time for the very last coupon is always the longest.
- $2H_2 - 1H_1 = 2$, $100H_{100} - 99H_{99} = 6.177$.
- $142H_{142} \approx 786.19$.
- $\mathbb{E}(X) = nH_n = n\log n + \gamma n + \frac{1}{2} + o(1)$.
- $\mathbb{E}(X)$ grows at the rate of $O(n\log(n))$, which is not too bad.
- $Var(X) < \frac{\pi^2}{6}n^2$, Chebyshev inequality gives:

$$\mathbb{P}(|X - nH_n| \geq cn) \leq \frac{\pi^2}{6c^2}. \tag{1}$$

Markov Chain (Omitted)

- Due to independence inherited from the equal probability assumption, we can construct a Markov Chain.
- Details omitted.
- State space is $S = \{0, 1, \dots, n\}$, and the transition probability matrix is:

$$P = \begin{pmatrix} 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1/n & (n-1)/n & 0 & \dots & \dots & 0 \\ 0 & 0 & 2/n & (n-2)/n & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & 0 & (n-1)/n & 1/n \\ 0 & \dots & \dots & \dots & 0 & 0 & 1 \end{pmatrix}.$$



Introduction

Single Collection
- Equal Probability Assumption
  - Basic Probability Solution
  - Markov Chain Solution (Omitted)
- Unequal Probability Assumption
  - Basic Probability Solution
  - Sensitivity of Overall Collection
  - Poisson Process Solution (Omitted)

Other Generalisations

Pokemon Problem Simulations

Single collection: unequal probability

- Now, let's assume that probability of collecting each coupon is **different**. i.e. p_i are distinct and $\sum_{i=1}^{n} p_i = 1$.
- If we **redefine** X_i to be the random number of coupons we need to buy to obtain the first coupon of type i, then each $X_i \sim Geo(p_i)$.
- $X = \max\{X_1, \ldots, X_n\}$ is the number of coupons we should buy till completion of the set.
- Any guess for the expected total waiting time $\mathbb{E}(X)$?

Single collection: unequal probability

- Now, let's assume that probability of collecting each coupon is **different**. i.e. p_i are distinct and $\sum_{i=1}^{n} p_i = 1$.
- If we **redefine** X_i to be the random number of coupons we need to buy to obtain the first coupon of type i, then each $X_i \sim Geo(p_i)$.
- $X = \max\{X_1, \ldots, X_n\}$ is the number of coupons we should buy till completion of the set.
- Any guess for the expected total waiting time $\mathbb{E}(X)$?

Solution via basic probability

$$\mathbb{E}(X) = \mathbb{E}\left[\max\{X_1, \ldots, X_n\}\right] \tag{2}$$

(3)

(4)

- The complication here is that X_i's are no longer independent

Solution via basic probability

$$\mathbb{E}(X) = \mathbb{E}\left[\max\{X_1, \ldots, X_n\}\right] \tag{2}$$

$$\tag{3}$$

$$\tag{4}$$

- The complication here is that X_i's are no longer independent
- $\min(X_i, X_j) \sim Geo(p_i + p_j)$, for $i \neq j$. This generalises to any finite number of X_i's.

Solution via basic probability

$$\mathbb{E}(X) = \mathbb{E}\left[\max\{X_1, \ldots, X_n\}\right] \tag{2}$$

(3)

(4)

- The complication here is that X_i's are no longer independent
- $\min(X_i, X_j) \sim Geo(p_i + p_j)$, for $i \neq j$. This generalises to any finite number of X_i's.
- The Maximum-Minimum Identity: $\mathbb{E}\left[\max(X_i, X_j))\right] = \mathbb{E}(X_i) + \mathbb{E}(X_j) - \mathbb{E}\left[\min(X_i, X_j)\right]$, the last summand is of the form $1/(p_i + p_j)$, which also generalises.

Solution via basic probability

$$\mathbb{E}(X) = \mathbb{E}\left[\max\{X_1, \ldots, X_n\}\right] \tag{2}$$

$$= \sum_{m=1}^{n} (-1)^{m-1} \sum_{1 \le j_1 < \cdots < j_m \le n} \frac{1}{p_{j_1} + \cdots + p_{j_m}} \tag{3}$$

$$\tag{4}$$

- The complication here is that X_i's are no longer independent
- $\min(X_i, X_j) \sim Geo(p_i + p_j)$, for $i \neq j$. This generalises to any finite number of X_i's.
- The Maximum-Minimum Identity: $\mathbb{E}\left[\max(X_i, X_j)\right] = \mathbb{E}(X_i) + \mathbb{E}(X_j) - \mathbb{E}\left[\min(X_i, X_j)\right]$, the last summand is of the form $1/(p_i + p_j)$, which also generalises.

Solution via basic probability

$$\mathbb{E}(X) = \mathbb{E}\left[\max\{X_1, \ldots, X_n\}\right] \tag{2}$$

$$= \sum_{m=1}^{n} (-1)^{m-1} \sum_{1 \le j_1 < \cdots < j_m \le n} \frac{1}{p_{j_1} + \cdots + p_{j_m}} \tag{3}$$

$$= \int_0^\infty \left(1 - \prod_{i=1}^{n} \left(1 - e^{-p_i x}\right)\right) \, dx. \tag{4}$$

- The complication here is that X_i's are no longer independent
- $\min(X_i, X_j) \sim Geo(p_i + p_j)$, for $i \neq j$. This generalises to any finite number of X_i's.
- The Maximum-Minimum Identity: $\mathbb{E}\left[\max(X_i, X_j))\right] = \mathbb{E}(X_i) + \mathbb{E}(X_j) - \mathbb{E}\left[\min(X_i, X_j)\right]$, the last summand is of the form $1/(p_i + p_j)$, which also generalises.

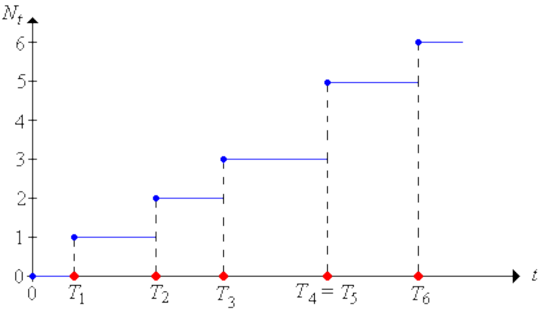
Some characteristics

- Relations to Stochastic Processes and Extreme Value Theory.
- Each prize in McDonald's Monopoly competition is exactly like this.
- The expression is not simple.
- Sensitive to introduction of low p_i's.
- $\mathbb{E}(X)_{(1/3,1/3,1/3)} = 5.5$
- $\mathbb{E}(X)_{(0.1,0.1,0.8)} = 15.03$
- $\mathbb{E}(X)_{(0.02,0.49,0.49)} = 50$
- Visualisation of the family $(p_1, p_2, p_3, \mathbb{E}(X))$.

Some characteristics

- Relations to Stochastic Processes and Extreme Value Theory.
- Each prize in McDonald's Monopoly competition is exactly like this.
- The expression is not simple.
- Sensitive to introduction of low p_i's.
- $\mathbb{E}(X)_{(1/3,1/3,1/3)} = 5.5$
- $\mathbb{E}(X)_{(0.1,0.1,0.8)} = 15.03$
- $\mathbb{E}(X)_{(0.02,0.49,0.49)} = 50$
- Visualisation of the family $(p_1, p_2, p_3, \mathbb{E}(X))$.

Poisson Process (Omitted)



- Avoids the horrible summations and integral tricks. The trade-off being you need to know something about Poisson Processes (STAT3911).
- Come to SUMS Lightening Talk in Week 13!

Introduction

Single Collection
- Equal Probability Assumption
  - Basic Probability Solution
  - Markov Chain Solution (Omitted)
- Unequal Probability Assumption
  - Basic Probability Solution
  - Sensitivity of Overall Collection
  - Poisson Process Solution (Omitted)

Other Generalisations

Pokemon Problem Simulations

Other Generalisations (Omitted)

- Consider $m-$collection problem now, which intuitively, should be less than m multiplied by their Single Collection counterpart.
- Equal Probability Solution:

$$\mathbb{E}(X) = n \int_0^\infty \left[1 - \left(1 - S_m(t)e^{-t}\right)^n\right] \, dt, \tag{5}$$

  where $S_m(t) = \sum_{k=0}^{m-1} \frac{t^k}{k!}$. For large m, this approaches mn.
- Unequal Probability Solution:
  - Bounds exist, see paper.
  - Markov Chain approach should work, but must require software computations.
- Other generalisations: arrival in batches of constant sizes.

Other Generalisations (Omitted)

- Consider $m-$collection problem now, which intuitively, should be less than m multiplied by their Single Collection counterpart.
- Equal Probability Solution:

$$\mathbb{E}(X) = n \int_0^\infty \left[1 - (1 - S_m(t)e^{-t})^n\right] \, dt, \tag{5}$$

where $S_m(t) = \sum_{k=0}^{m-1} \frac{t^k}{k!}$. For large m, this approaches mn.

- Unequal Probability Solution:
  - Bounds exist, see paper.
  - Markov Chain approach should work, but must require software computations.
- Other generalisations: arrival in batches of constant sizes.

Other Generalisations (Omitted)

- Consider $m-$collection problem now, which intuitively, should be less than m multiplied by their Single Collection counterpart.
- Equal Probability Solution:

$$\mathbb{E}(X) = n \int_0^\infty \left[1 - (1 - S_m(t)e^{-t})^n \right] \, dt, \tag{5}$$

where $S_m(t) = \sum_{k=0}^{m-1} \frac{t^k}{k!}$. For large m, this approaches mn.
- Unequal Probability Solution:
  - Bounds exist, see paper.
  - Markov Chain approach should work, but must require software computations.
- Other generalisations: arrival in batches of constant sizes.

Other Generalisations (Omitted)

- Consider $m-$collection problem now, which intuitively, should be less than m multiplied by their Single Collection counterpart.
- Equal Probability Solution:

$$\mathbb{E}(X) = n \int_0^\infty \left[1 - (1 - S_m(t)e^{-t})^n\right] \, dt, \tag{5}$$

  where $S_m(t) = \sum_{k=0}^{m-1} \frac{t^k}{k!}$. For large m, this approaches mn.
- Unequal Probability Solution:
  - Bounds exist, see paper.
  - Markov Chain approach should work, but must require software computations.
- Other generalisations: arrival in batches of constant sizes.

The Pokemon Collector's Problem

- Pokemon problem has:
  - unequal probabilities
  - multiple collection with differing termination point (evolution)
  - arrival in batches of different sizes (eggs)
- A theoretical closed form does not exist in literature so far.
- I am sure if you tweak features in this model here and there, you can potentially get a paper out of this and a job.

The Pokemon Collector's Problem

- Pokemon problem has:
  - unequal probabilities
  - multiple collection with differing termination point (evolution)
  - arrival in batches of different sizes (eggs)
- A theoretical closed form does not exist in literature so far.
- I am sure if you tweak features in this model here and there, you can potentially get a paper out of this and a job.

Introduction

Single Collection
- Equal Probability Assumption
  - Basic Probability Solution
  - Markov Chain Solution (Omitted)
- Unequal Probability Assumption
  - Basic Probability Solution
  - Sensitivity of Overall Collection
  - Poisson Process Solution (Omitted)

Other Generalisations

Pokemon Problem Simulations

Real-world considerations

- The game is even more complicated, with eggs, evolutions and human irrationality.
- Recent updates may have fundamentally changed rate parameters. The main challenge of our simulation is to get some sensible rates and see how PP compares to classical CCP models.
- Capture completion is highly **sensitive** to existence of low rates.
- Data trustworthiness, untidiness and missingness.

Real-world considerations

- The game is even more complicated, with eggs, evolutions and human irrationality.
- Recent updates may have fundamentally changed rate parameters. The main challenge of our simulation is to get some sensible rates and see how PP compares to classical CCP models.
- Capture completion is highly **sensitive** to existence of low rates.
- Data trustworthiness, untidiness and missingness.

Real-world considerations

- The game is even more complicated, with eggs, evolutions and human irrationality.
- Recent updates may have fundamentally changed rate parameters. The main challenge of our simulation is to get some sensible rates and see how PP compares to classical CCP models.
- Capture completion is highly **sensitive** to existence of low rates.
- Data trustworthiness, untidiness and missingness.

Step 1: Estimate rates

> In God we trust. All others bring data —
>
> William Edwards Deming

- While someone published a sensible table of capture rates. However, simulations suggested these were underestimated.
- By running a very close simulation, this rate was transformed to get more sensible expected number of captures. This sets the benchmark around 4,700 Pokemons.
- Again, let me assure you this is not easy.
- See demonstration.

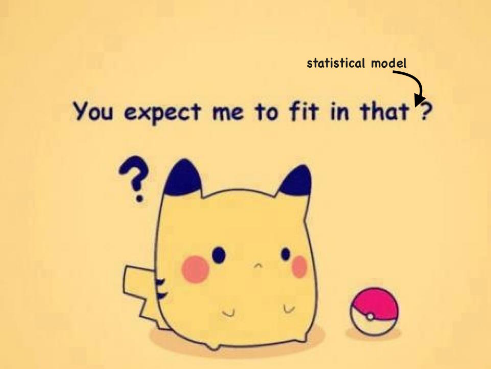
Step 1: Estimate rates

> In God we trust. All others bring data —
>
> William Edwards Deming

- While someone published a sensible table of capture rates. However, simulations suggested these were underestimated.
- By running a very close simulation, this rate was transformed to get more sensible expected number of captures. This sets the benchmark around 4,700 Pokemons.
- Again, let me assure you this is not easy.
- See demonstration.

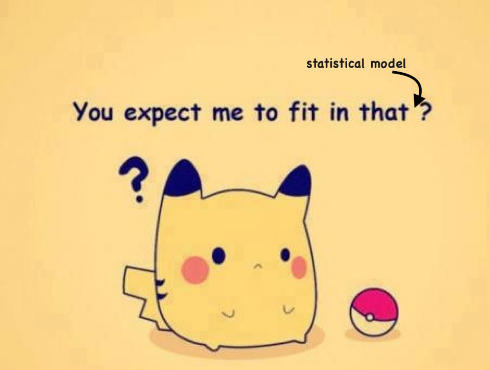
Step 2: Compare this to other strategies

- Now, assume we can **ONLY** access higher level Pokemons via evolution of lower level Pokemons. This on average, require roughly 10,310 Pokemons, taking into account specific number of collections.
- Compare this to the benchmark of 4,700 in a simulation where capture of higher level Pokemons is possible.



Step 2: Compare this to other strategies

- Now, assume we can **ONLY** access higher level Pokemons via evolution of lower level Pokemons. This on average, require roughly 10,310 Pokemons, taking into account specific number of collections.
- Compare this to the benchmark of 4,700 in a simulation where capture of higher level Pokemons is possible.



Trivial lessons

Rates for higher stage Pokemons are in a sweet spot — non-zero capture probability made sure you don't end up playing the game via evolution for years.

Theorem

Skipping research to catch a Charizard is therefore a high return investment in your future.

Trivial lessons

Rates for higher stage Pokemons are in a sweet spot — non-zero capture probability made sure you don't end up playing the game via evolution for years.

Theorem

Skipping research to catch a Charizard is therefore a high return investment in your future.

Concluding remarks

- Theory sometimes only go so far, which leave room for explorations.
- Theory and application are not mortal enemies.
- Mathematics and Statistics are not and should not be dreadful subjects. They exist outside the realm of tutorial sheets and assignments.
- Worst case: Mathematics and Statistics are prosecutors of our "intuitions" and "common sense".

References

- *The Coupon Collector's Problem*, M. Ferrante, M. Saltalamacchia, (2014)
- *Introduction to Probability Models.* S.Ross.
- *A First Course in Probability.* S. Ross.
- *STAT3911: Stochastic Processes Lecture Notes.* R. Kawaii.