The Birthday Problem (2.7)

Prof. Tesler

Math 186
January 21, 2009

Fun Party Fact

In a group of 23 or more randomly chosen people, there is at least a 50% chance that at least two of them share the same birthday.

General Setup

- n days in a year.
- k people.
- Birthdays are uniform (each person has probability $1/n$ for each possible day) and birthdays of different people are independent:
  - If your club has a party for everyone with a January birthday, the people with January birthdays may be over-represented.
  - In a club for twins, the birthdays also would not be independent.
- What's the probability p that at least two people share a birthday? Equivalently, compute $q = 1 - p$, the probability that all birthdays are different.

Probability all birthdays are different

Example: 3 people

- First person has a unique birthday with probability $\frac{n}{n} = 1$.
- Second person has a birthday different from the first with probability $\frac{n-1}{n}$.
- Given that the first two birthdays were different, the third person has a birthday different from those with probability $\frac{n-2}{n}$.
- $q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n}$

General case

$$q = \prod_{r=1}^{k} P\left(r\text{th birthday different from 1st } r-1 \;\middle|\; \text{1st } r-1 \text{ distinct}\right)$$

$$= \prod_{r=1}^{k} \frac{n-r+1}{n} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}$$

Probability all birthdays are different, 2nd derivation

- The sample space is all k-tuples of integers $1, \ldots, n$:

$$S = \{\, (x_1, x_2, \ldots, x_k) \;:\; 1 \leqslant x_i \leqslant n \,\}$$

  where the ith person has birthday x_i. Note $N(S) = n^k$.
- E.g., number the days of the year $1, 2, \ldots, 365$.
  $(33, 2, 365)$ means the first person is born the 33rd day of the year (Feb. 2), the second is born Jan. 2, the third is born Dec. 31.
- Let A be the event that all birthdays are different.
- $N(A) = {}_nP_k = n(n-1)(n-2)\ldots(n-k+1)$
- $P(A) = N(A)/N(S) = \frac{{}_nP_k}{n^k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}$

Probability all birthdays are different, approximation

We will also give an approximate formula for q:

$$q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} \qquad \approx \exp\left(-\frac{k^2}{2n}\right) \quad \text{for } k \ll n.$$

Question

How large a group of people is needed for at least a 90% chance that at least two share a birthday?

Answer

- $p \geqslant 90\%$ gives $q = 1 - p \leqslant 10\%$.
- We could chug away the exact equation $q = \frac{365}{365}\frac{364}{365} \cdots \frac{366-k}{365}$ on a calculator for $k = 1, 2, 3, \ldots$ until we get $q < 10\%$.
- Or we can solve for k from the approximate formula:

$$q \approx \exp(-\frac{k^2}{2n}) \quad \ln(q) \approx -\frac{k^2}{2n} \quad k \approx +\sqrt{-2n\ln(q)} = +\sqrt{-2n\ln(1-p)}$$

- Note $1 - p < 1$ so $\ln(1-p) < 0$ and $-2n\ln(1-p) > 0$.

Probability all birthdays are different, approximation

$$q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} \qquad \approx \exp\left(-\frac{k^2}{2n}\right) \quad \text{for } k \ll n.$$

- For at least a 90% chance that two people share a birthday, use $k = 41$:

k	q with exact formula	q with approx formula
40	0.1087	0.1117
41	0.0968	0.0999

- How about for $p = 50\%$?

Party problem

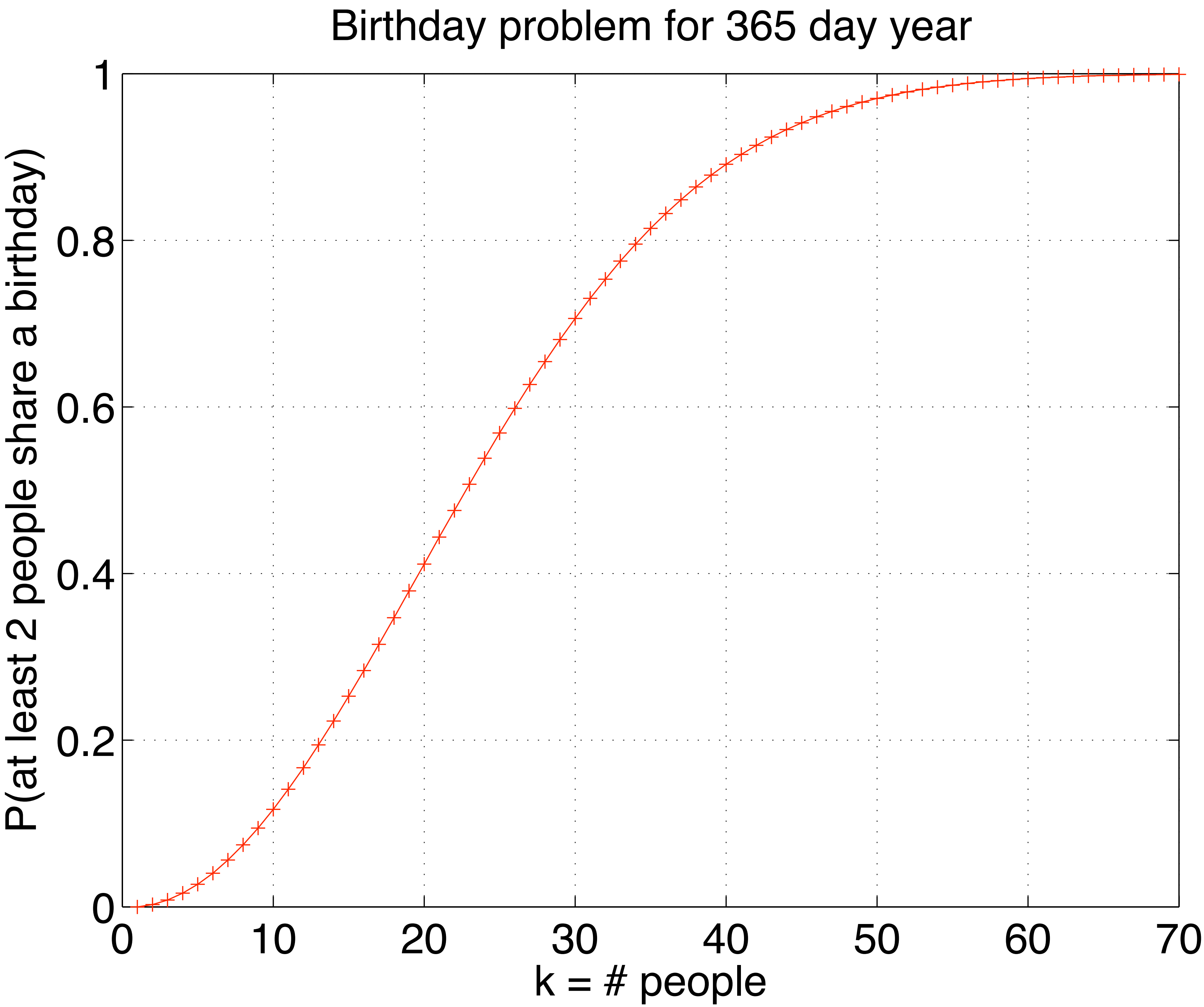
- $q = 1 - p = .50$ and $k \approx \sqrt{-2(365)\ln(.50)} = 22.49$
- In a group of 23 randomly selected people, there's a $p \approx 1 - \exp(-\frac{23^2}{2(365)}) = 51.55\%$ chance that two share a birthday. (The exact formula gives $p = 1 - \frac{365}{365}\frac{364}{365}\cdots\frac{343}{365} \approx 50.73\%$.)
- In a group of 23 or more randomly selected people, there's at least a 50% chance that two share a birthday.

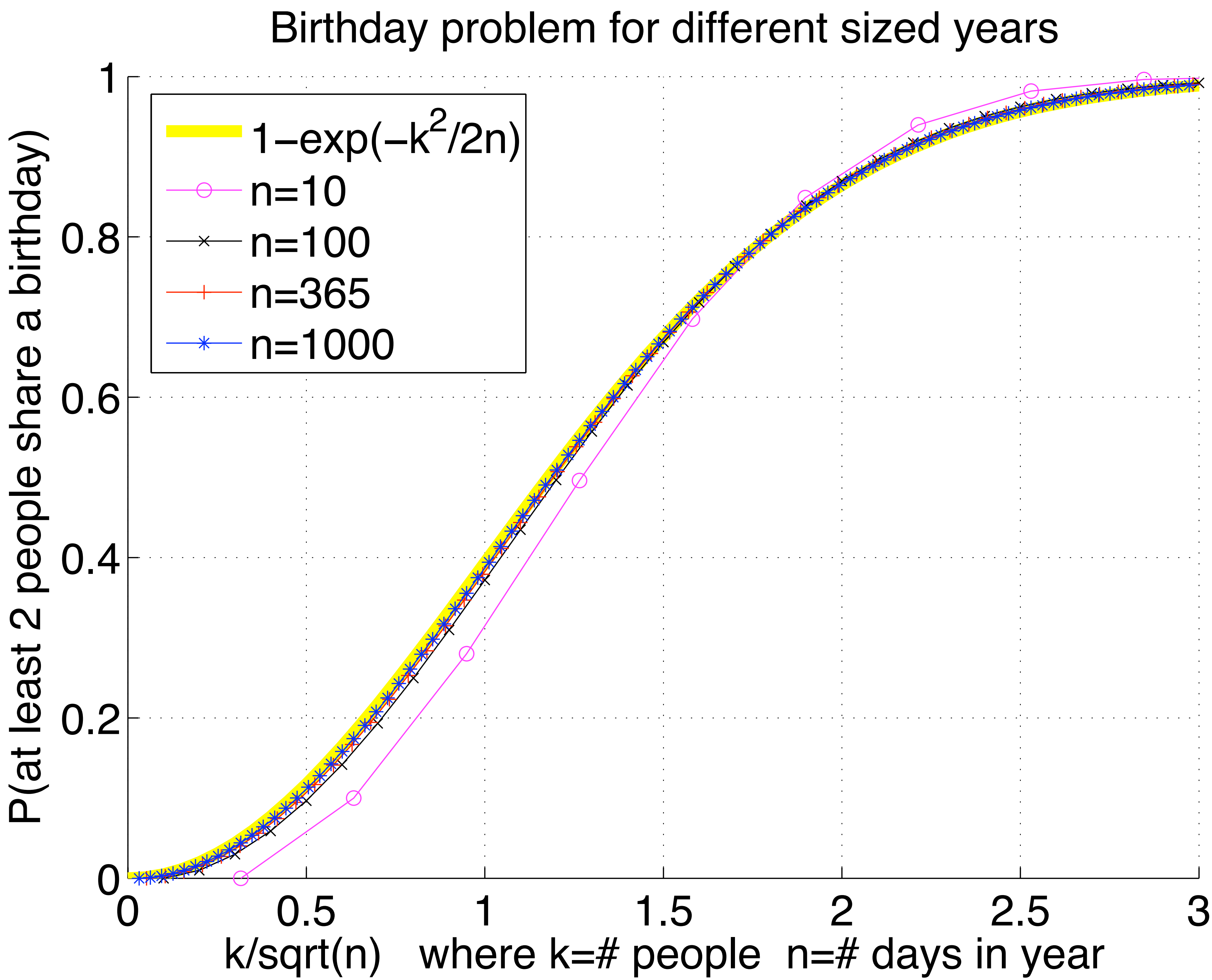
Varying the number of days in a year

Using $k \approx \sqrt{-2\ln(1-p)}\sqrt{n}$ gives

p	**k in n day year**	**k in 365 day year**
$.5$	$1.18\sqrt{n}$	23
$.7$	$1.55\sqrt{n}$	30
$.9$	$2.15\sqrt{n}$	41
$.99$	$3.03\sqrt{n}$	58

On the graphs that follow, we plot the exact probability formula.
On the second graph, we also superimpose the approximate probability formula in yellow.


Birthday problem for 365 day year
P(at least 2 people share a birthday)
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
k = # people


Birthday problem for different sized years
1−exp(−k^2/2n)
n=10
n=100
n=365
n=1000
P(at least 2 people share a birthday)
0
0.2
0.4
0.6
0.8
1
0
0.5
1
1.5
2
2.5
3
k/sqrt(n)   where k=# people n=# days in year

Derivation of approximation formula

- Start from the exact formula
$$q = \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n}$$

- Take the logarithm to convert the product to a sum:
$$\ln(q) = \ln\left(\frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n}\right) = \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right)$$

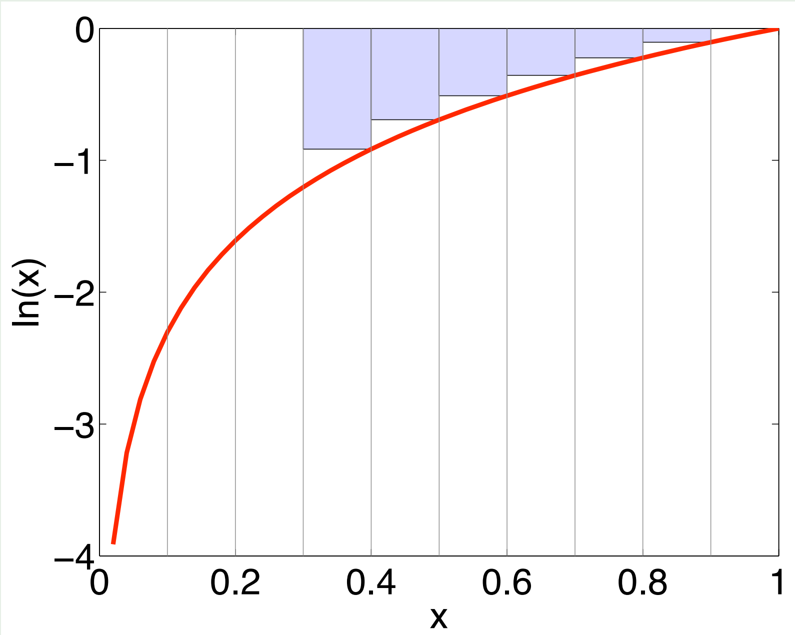
- **Trick:** Multiply by $1 = n \cdot \frac{1}{n}$ and approximate it as an integral:
$$\ln(q) = n \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right) \frac{1}{n} \approx n \int_{1-k/n}^{1} \ln(x)\,dx$$
Note: bounds are $\frac{n-k}{n} = 1 - \frac{k}{n}$ *and* $\frac{n}{n} = 1$

Derivation of approximation formula

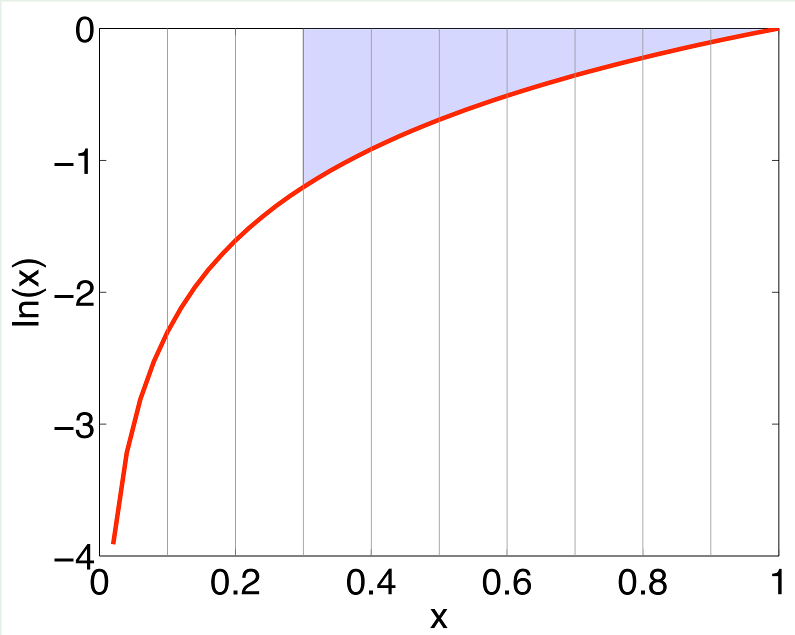
$$\ln(q) = n \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right)\frac{1}{n} \approx n \int_{1-k/n}^{1} \ln(x)\,dx$$

Example: $n = 10$, $k = 7$; sum is negative area indicated

$$\sum_{r=4}^{10} \ln\left(\tfrac{r}{10}\right)\tfrac{1}{10}$$



$$\int_{.4}^{1} \ln(x)\,dx$$



Derivation of approximation formula

$$\ln(q) \approx n \int_{1-k/n}^{1} \ln(x)\,dx = n\Big(x\big(\ln(x)-1\big)\Big)\Big|_{1-k/n}^{1}$$
$$= n\Big(1\big(\ln(1)-1\big) - (1-k/n)\big(\ln(1-k/n)-1\big)\Big)$$
$$= n\Big(-k/n - (1-k/n)\big(\ln(1-k/n)\big)\Big)$$

- Using the Taylor series $\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots$ gives
$$(1-x)\ln(1-x) = -x + \frac{x^2}{2\cdot 1} + \frac{x^3}{3\cdot 2} + \frac{x^4}{4\cdot 3} + \cdots$$
- Use this (with $x = k/n$) and plug into the approximation for $\ln(q)$. The leading term is
$$\ln(q) \approx n\Big(-\frac{k}{n} + \frac{k}{n} - \frac{k^2}{2\cdot 1\cdot n^2} - \frac{k^3}{3\cdot 2n^3} - \frac{k^4}{4\cdot 3n^4} - \cdots\Big) \approx -\frac{k^2}{2n}\,.$$
so $p = 1 - q \approx 1 - \exp\left(-\frac{k^2}{2n}\right)$.
- The graphs show this approximation is pretty good except for small n. It's possible to quantify the error analytically also.