Math 180A, Fall 2005, Prof. Tesler – October 3, 2005
The Birthday Problem

If there are n days in the year and k people, and if each person has probability $1/n$ of being born on each day independently of the other people, then the probability that at least two of them share a birthday is

$$b(n,k) = 1 - \frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n}$$

We want to answer questions such as "in a 365 day year, how many people are necessary to have a 90% probability that at least two people share a birthday?" If you are told a specific number of days n and a specific probability p (here, $n = 365$ and $p = .90$), you can chug away on your calculator with $k = 1, 2, \ldots$ until you hit the answer. We want to solve this in a more general way, leaving n and p as variables: given n and p, find k that makes $b(n,k) \approx p$.

Below we will show that $b(n,k) \approx 1 - \exp(-\frac{k^2}{2n})$, provided k is "small" compared to n. Solving $1 - \exp(-\frac{k^2}{2n}) = p$ gives $\boxed{\boldsymbol{k \approx \sqrt{-2\ln(1-p)}\sqrt{n}}}$. (Note: $1 - p < 1$ so $\ln(1-p) < 0$ and $-2\ln(1-p) > 0$.) Graphs of $b(n,k)$ are shown on the next page for different n's. This gives some approximations as follows:

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

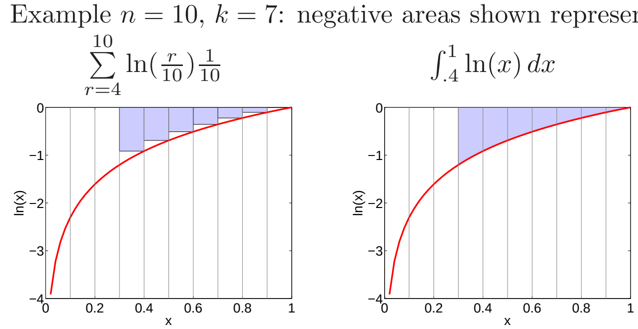
Here's how to derive the approximation formula for $b(n,k)$. Convert the product into a sum by taking a logarithm:

$$S = \ln\left(\frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n}\right) = \sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right) .$$

So $b(n,k) = 1 - e^S$. This sum can be approximated using calculus. Multiply and divide by $1/n$ to turn it into the "right-hand Riemann sum" approximating the integral shown:

$$\begin{aligned} S &= \frac{\sum_{r=n-k+1}^{n} \ln\left(\frac{r}{n}\right)\frac{1}{n}}{1/n} \\ &\approx n \int_{1-k/n}^{1} \ln(x)\,dx \\ &\qquad \text{(Note: bounds are } \tfrac{n-k}{n} = 1 - \tfrac{k}{n} \text{ and } \tfrac{n}{n} = 1) \\ &= n\Big(x\big(\ln(x) - 1\big)\Big)\Big|_{1-k/n}^{1} \\ &= n\Big(1\big(\ln(1) - 1\big) - \big(1 - k/n\big)\big(\ln(1-k/n) - 1\big)\Big) \\ &= n\Big(-k/n - (1-k/n)\big(\ln(1-k/n)\big)\Big) \end{aligned}$$

Example $n = 10$, $k = 7$: negative areas shown represent



Use these Taylor series:

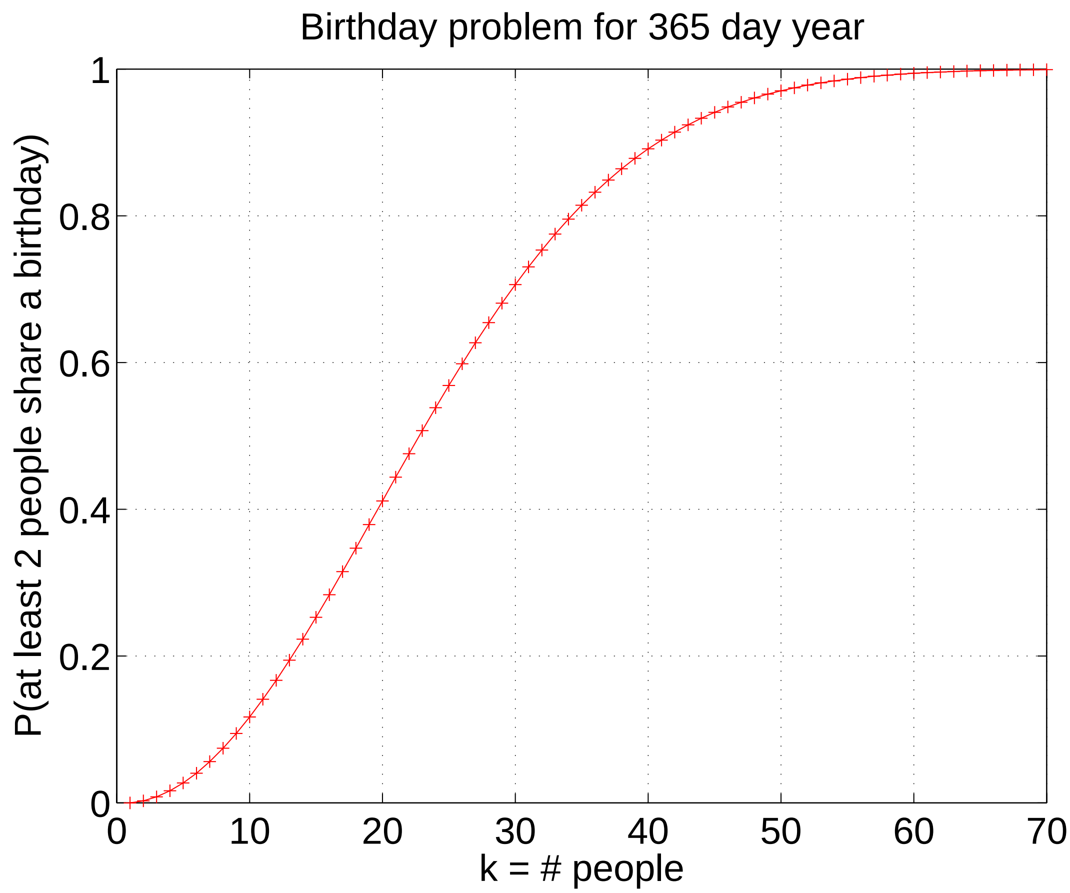
$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \frac{x^5}{5} - \cdots \qquad (1-x)\ln(1-x) = -x + \frac{x^2}{2\cdot 1} + \frac{x^3}{3\cdot 2} + \frac{x^4}{4\cdot 3} + \frac{x^5}{5\cdot 4} + \cdots$$

Plug the second Taylor series into the approximation for S, and keep only the leading term, to obtain

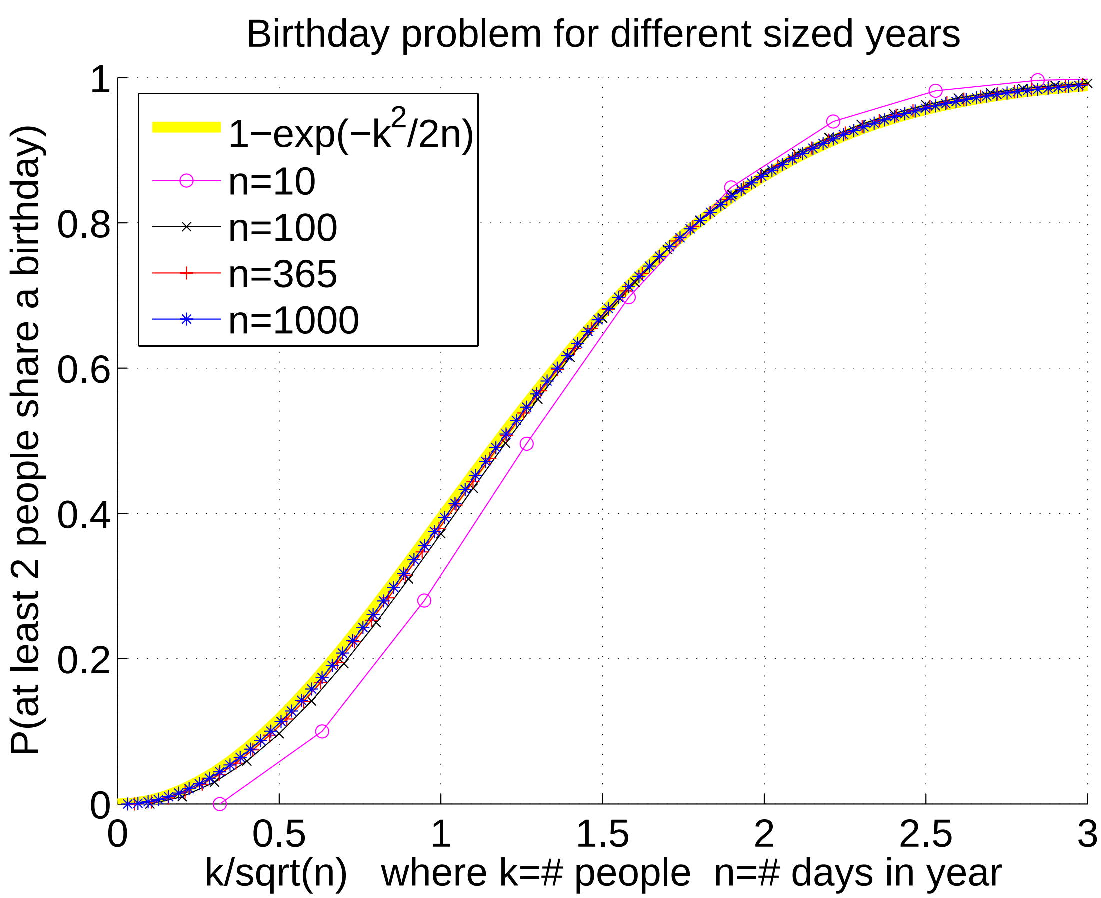
$$S \approx n\left(-\frac{k}{n} + \frac{k}{n} - \frac{k^2}{2\cdot 1\cdot n^2} - \frac{k^3}{3\cdot 2n^3} - \frac{k^4}{4\cdot 3n^4} - \cdots\right) \approx -\frac{k^2}{2n} .$$

Plug this back in to $b(n,k) = 1 - e^S$ to obtain $b(n,k) \approx 1 - \exp(-\frac{k^2}{2n})$.

Each of the approximations above could be done with more details using the techniques of the Math 20 series to quantify just how good these estimates are. However, the graphs on the next page show empirically that these estimates are pretty good except for very small n.



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



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