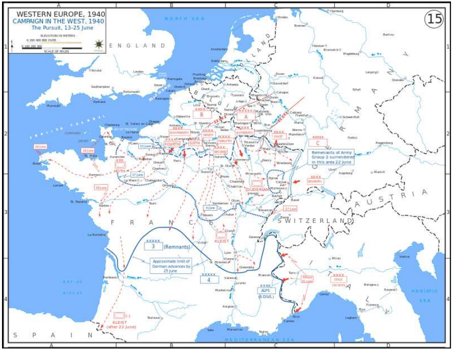
The German Tank Problem: Math/Stats at War!

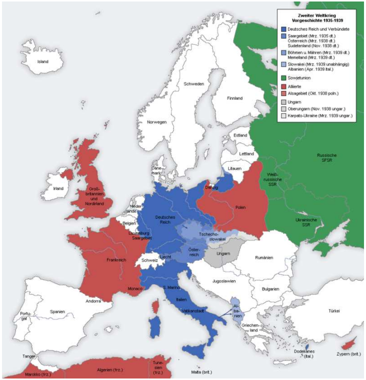


Steven J. Miller: Williams College
(sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu)
http://www.williams.edu/Mathematics/sjmiller

UNC Greensboro REU, June 9, 2022

Introduction

The German Tank Problem



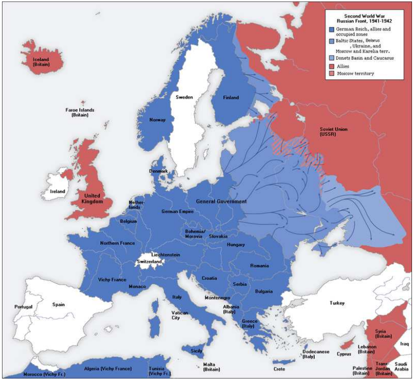


Figure: Left: Europe before the start of major hostilities. Right: Europe in 1942. Wikimedia Commons (author San Jose).
Video: https://www.youtube.com/watch?v=WOVEy1tC7nk

Approaches

Figure: German tank production. Photographer Rudolf Schmidt, provided to Wikimedia Commons by the German Federal Archive.

How to find the number of tanks?

Approaches

Figure: German tank production. Photographer Rudolf Schmidt, provided to Wikimedia Commons by the German Federal Archive.

How to find the number of tanks? Spies vs Math/Stats!

Related Problems: My Experience

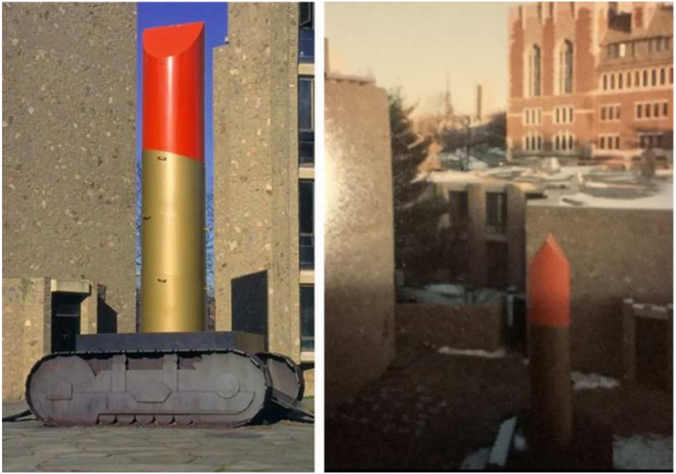


Figure: Morse tank problem. R: SJMiller, College Years. L: https://www.pinterest.com/pin/288230444874671007/.

The German Tank Problem: General Statement

German Tank Problem

After a battle see serial numbers $s_1, s_2, \dots, s_k$ on captured and destroyed tanks; how many tanks were produced?

The German Tank Problem: Mathematical Formulation

Original formulation: Tanks numbered from 1 to N, observe k, largest seen is m, estimate N with $\widehat{N} = \widehat{N}(m, k)$.

New version: Tanks numbered from N_1 to N_2, observe k, smallest seen is m_1, largest m_2 (so spread $s = m_2 - m_1$), estimate $N = N_2 - N_1 + 1$ with $\widehat{N} = \widehat{N}(s, k)$.

Importance

Dangers of under-estimating and over-estimating.

Importance

Dangers of under-estimating and over-estimating.

President Lincoln: *If General McClellan isn't going to use his army, I'd like to borrow it for a time.*

Importance

Dangers of under-estimating and over-estimating.

President Lincoln: *If General McClellan isn't going to use his army, I'd like to borrow it for a time.*

From Wikipedia: Although outnumbered two-to-one, Lee committed his entire force, while McClellan sent in less than three-quarters of his army, enabling Lee to fight the Federals to a standstill. McClellan's persistent but erroneous belief that he was outnumbered contributed to his cautiousness throughout the campaign.

Outline of the Talk

- Mathematical Preliminaries (binomial coefficients).
- GTP: Original Formulation (start at 1).
- GTP: New Version (unknown start).
- Regression and Conjecturing.

Mathematical Preliminaries

Basic Combinatorics

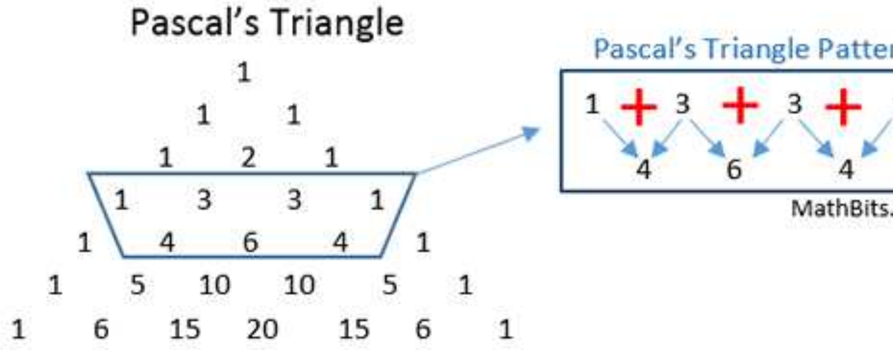
- $n!$: Number of ways to order n objects, equals $n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1$, with $0! = 1$.
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$: Number of ways to choose k items from n items when order doesn't matter.
- Binomial Theorem:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Pascal's Identity

Pascal's identity

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}.$$



Pascal's Identity

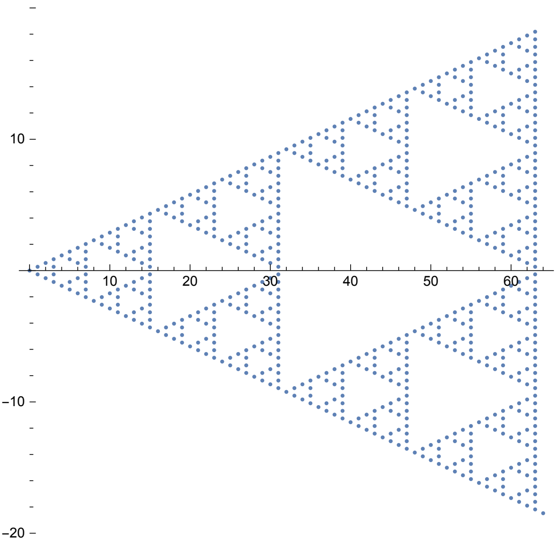
Pascal's identity

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}.$$

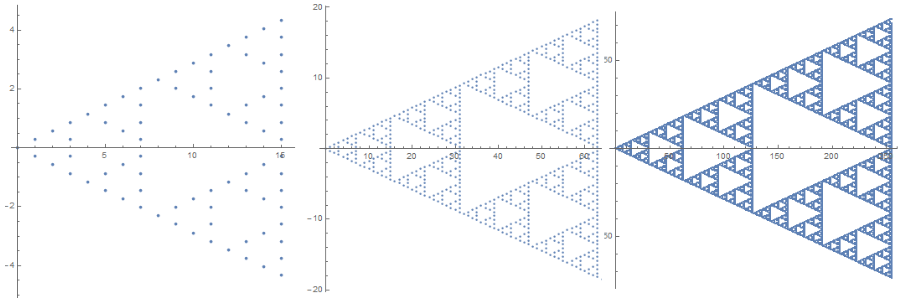
Proof: Assume n Red Sox fans, 1 Yankee fan, how many ways to choose a group of k?

$$\binom{n+1}{k} = \binom{1}{0}\binom{n}{k} + \binom{1}{1}\binom{n}{k-1}.$$

Pascal's Triangle Mod 2

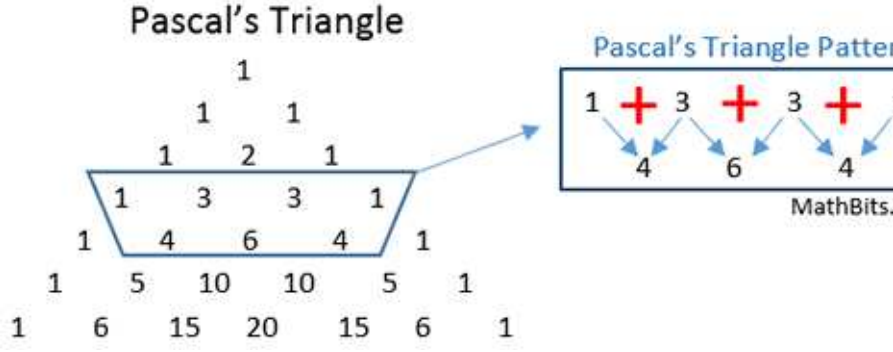


Pascal's Triangle Mod 2



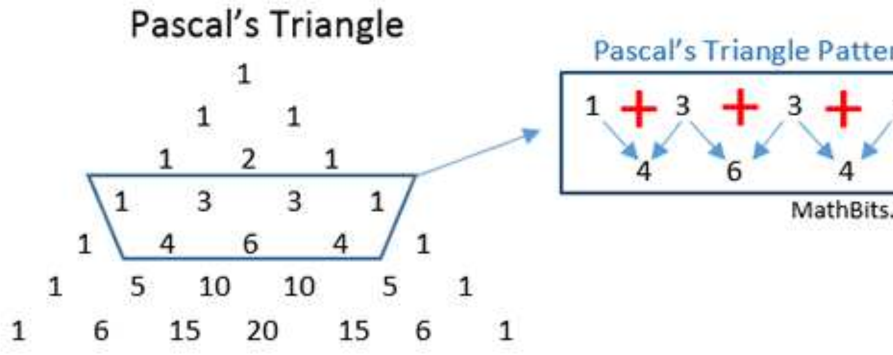
https://www.youtube.com/watch?v=tt4_4YajqRM

Pascal's Triangle



Needed Identity

$$\sum_{m=k}^{N} \binom{m}{k} = \binom{N+1}{k+1}.$$



Needed Identity

$$\sum_{m=k}^{N} \binom{m}{k} = \binom{N+1}{k+1}.$$

Proof: Induct on N, noting that k is fixed.

Needed Identity

$$\sum_{m=k}^{N} \binom{m}{k} = \binom{N+1}{k+1}.$$

Proof: Induct on N, noting that k is fixed.

The base case ($N = k$) is readily established.

Needed Identity

$$\sum_{m=k}^{N} \binom{m}{k} = \binom{N+1}{k+1}.$$

Proof: Induct on N, noting that k is fixed.

The base case ($N = k$) is readily established.
Inductive step:

$$\begin{aligned} \sum_{m=k}^{N+1} \binom{m}{k} &= \left(\sum_{m=k}^{N} \binom{m}{k} \right) + \binom{N+1}{k} \\ &= \binom{N+1}{k+1} + \binom{N+1}{k} = \binom{N+2}{k+1}, \end{aligned}$$

where the last equality follows from Pascal's identity. □

GTP: Original

Conjecturing the Formula

How should $\widehat{N}$ depend on m (maximum tank number observed) and k (number of tanks observed)?

Conjecturing the Formula

How should $\widehat{N}$ depend on m (maximum tank number observed) and k (number of tanks observed)?

Expect $\widehat{N} = m + g(m, k)$, simplest guess $g(m, k) = bm/k$ for some constant b.

Conjecturing the Formula

How should $\widehat{N}$ depend on m (maximum tank number observed) and k (number of tanks observed)?

Expect $\widehat{N} = m + g(m, k)$, simplest guess $g(m, k) = bm/k$ for some constant b.

Answer:

$$\widehat{N} = m\left(1 + \frac{1}{k}\right) - 1;$$

is this reasonable?

Conjecturing the Formula

How should $\widehat{N}$ depend on m (maximum tank number observed) and k (number of tanks observed)?

Expect $\widehat{N} = m + g(m, k)$, simplest guess $g(m, k) = bm/k$ for some constant b.

Answer:

$$\widehat{N} = m\left(1 + \frac{1}{k}\right) - 1;$$

is this reasonable?

Note sanity checks at $k = 1$ and $k = N$.

Proving the Formula: Step 1: Sample Maximum Probability

Lemma

Let M be the random variable for the maximum number observed, and let m be the value we see. There is zero probability of observing a value smaller than k or larger than N, and for $k \leq m \leq N$

$$\text{Prob}(M = m) \ = \ \frac{\binom{m}{k} - \binom{m-1}{k}}{\binom{N}{k}} \ = \ \frac{\binom{m-1}{k-1}}{\binom{N}{k}}.$$

Proving the Formula: Step 1: Sample Maximum Probability

Lemma

Let M be the random variable for the maximum number observed, and let m be the value we see. There is zero probability of observing a value smaller than k or larger than N, and for $k \leq m \leq N$

$$\mathrm{Prob}(M = m) \ = \ \frac{\binom{m}{k} - \binom{m-1}{k}}{\binom{N}{k}} \ = \ \frac{\binom{m-1}{k-1}}{\binom{N}{k}}.$$

Proof: If the largest is m then we have to choose that serial number, and now we must choose $k-1$ tanks from the $m-1$ smaller values; thus we find the probability is just $\binom{m-1}{k-1}/\binom{N}{k}$. □

Step 2: Expected Value Preliminaries

Definition of Expected Value:

$$\mathbb{E}[M] := \sum_{m=k}^{N} m \cdot \mathrm{Prob}(M = m).$$

Substituting:

$$\mathbb{E}[M] = \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}}.$$

Step 2: Computing $\mathbb{E}[M]$

$$\begin{aligned}
\mathbb{E}[M] &= \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}} \\
&= \sum_{m=k}^{N} m \frac{(m-1)!}{(k-1)!(m-k)!} \frac{k!(N-k)!}{N!} \\
&= \sum_{m=k}^{N} \frac{m!k}{k!(m-k)!} \frac{k!(N-k)!}{N!} \\
&= \frac{k \cdot k!(N-k)!}{N!} \sum_{m=k}^{N} \binom{m}{k} \\
&= \frac{k \cdot k!(N-k)!}{N!} \binom{N+1}{k+1}
\end{aligned}$$

Step 2: Computing $\mathbb{E}[M]$

$$
\begin{aligned}
\mathbb{E}[M] &= \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}} \\
&= \sum_{m=k}^{N} m \frac{(m-1)!}{(k-1)!(m-k)!} \frac{k!(N-k)!}{N!} \\
&= \sum_{m=k}^{N} \frac{m!k}{k!(m-k)!} \frac{k!(N-k)!}{N!} \\
&= \frac{k \cdot k!(N-k)!}{N!} \sum_{m=k}^{N} \binom{m}{k} \\
&= \frac{k \cdot k!(N-k)!}{N!} \frac{(N+1)!}{(k+1)!(N-k)!}
\end{aligned}
$$

Step 2: Computing $\mathbb{E}[M]$

$$
\begin{aligned}
\mathbb{E}[M] &= \sum_{m=k}^{N} m \frac{\binom{m-1}{k-1}}{\binom{N}{k}} \\
&= \sum_{m=k}^{N} m \frac{(m-1)!}{(k-1)!(m-k)!} \frac{k!(N-k)!}{N!} \\
&= \sum_{m=k}^{N} \frac{m!k}{k!(m-k)!} \frac{k!(N-k)!}{N!} \\
&= \frac{k \cdot k!(N-k)!}{N!} \sum_{m=k}^{N} \binom{m}{k} \\
&= \frac{k(N+1)}{k+1}.
\end{aligned}
$$

Step 3: Finding $\widehat{N}$

Showed $\mathbb{E}[M] = \frac{k(N+1)}{k+1}$, solve for N:

$$N \ = \ \mathbb{E}[M]\left(1 + \frac{1}{k}\right) - 1.$$

Substitute m (observed value for M) as best guess for $\mathbb{E}[M]$, we obtain our estimate for the number of tanks produced:

$$\widehat{N} \ = \ m\left(1 + \frac{1}{k}\right) - 1,$$

completing the proof. □

GTP: New

Conjecturing the Formula

How should $\widehat{N}$ depend on m_1, m_2 (min/max observed) and k (number of tanks observed)?

Let $N = N_2 - N_1 + 1$, $s = m_2 - m_1$.

Conjecturing the Formula

How should $\widehat{N}$ depend on m_1, m_2 (min/max observed) and k (number of tanks observed)?

Let $N = N_2 - N_1 + 1$, $s = m_2 - m_1$.

Expect $\widehat{N} = s + g(s, k)$, simplest guess $g(m, k) = bs/(k-1)$ for some constant b.

Conjecturing the Formula

How should $\widehat{N}$ depend on m_1, m_2 (min/max observed) and k (number of tanks observed)?

Let $N = N_2 - N_1 + 1$, $s = m_2 - m_1$.

Expect $\widehat{N} = s + g(s, k)$, simplest guess $g(m, k) = bs/(k-1)$ for some constant b.

Answer:

$$\widehat{N} = s\left(1 + \frac{2}{k-1}\right) - 1;$$

note sanity checks at $k = 2$ and $k = N$.

Proving the Formula: Step 1: Sample Maximum Probability

Lemma

Probability is 0 unless $k-1 \leq s \leq N_2 - N_1$, then

$$\mathrm{Prob}(S=s) \ = \ \frac{\sum_{m=N_1}^{N_2-s} \binom{s-1}{k-2}}{\binom{N_2-N_1+1}{k}} \ = \ \frac{(N-s)\binom{s-1}{k-2}}{\binom{N}{k}}.$$

Proof: Spread s must be at least $k-1$ (as we have k observations), and cannot be larger than $N_2 - N_1$.

If spread of s and smallest observed is m then largest is $m+s$. Choose exactly $k-2$ of the $s-1$ numbers in $\{m+1, m+2, \dots, m+s-1\}$. Have $\binom{s-1}{k-2}$ ways to do so, all the summands are the same. □

Outline of Proof

Rest of proof is similar to before, but more involved algebra.

Repeatedly multiply by 1 or add 0 to facilitate applications of binomial identities.

See appendix for details.

Comparison of Spies versus Statistics

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	
June 1941	244	1,550	
August 1942	327	1,550	

Figure: Comparison of estimates from statistics and spies to the true values. Table from https://en.wikipedia.org/wiki/German_tank_problem.

Comparison of Spies versus Statistics

Month	Statistical estimate	Intelligence estimate	German records
June 1940	169	1,000	122
June 1941	244	1,550	271
August 1942	327	1,550	342

Figure: Comparison of estimates from statistics and spies to the true values. Table from https://en.wikipedia.org/wiki/German_tank_problem.

Regression

See https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/MethodLeastSquares.pdf

Overview

Idea is to find *best-fit* parameters: choices that minimize error in a conjectured relationship.

Say observe y_i with input x_i, believe $y_i = ax_i + b$. Three choices:

$$\begin{aligned} E_1(a,b) &= \sum_{n=1}^{N} (y_i - (ax_i + b)) \\ E_2(a,b) &= \sum_{n=1}^{N} |y_i - (ax_i + b)| \\ E_3(a,b) &= \sum_{n=1}^{N} (y_i - (ax_i + b))^2 . \end{aligned}$$

Overview

Idea is to find *best-fit* parameters: choices that minimize error in a conjectured relationship.

Say observe y_i with input x_i, believe $y_i = ax_i + b$. Three choices:

$$\begin{aligned} E_1(a,b) &= \sum_{n=1}^{N} (y_i - (ax_i + b)) \\ E_2(a,b) &= \sum_{n=1}^{N} |y_i - (ax_i + b)| \\ E_3(a,b) &= \sum_{n=1}^{N} (y_i - (ax_i + b))^2 . \end{aligned}$$

Use sum of squares as calculus available.

Linear Regression

Explicit formula for values of a, b minimizing error $E_3(a, b)$. From

$$\partial E_3(a,b)/\partial a \;=\; \partial E_3(a,b)/\partial b \;=\; 0 :$$

$$\begin{aligned}
E_3(a,b) &= \sum_{n=1}^{N} (y_i - ax_i - b)^2 \\
\frac{\partial E_3(a,b)}{\partial a} &= \sum_{n=1}^{N} 2\,(y_i - ax_i - b)^{2-1} \cdot (-x_i) \\
\frac{\partial E_3(a,b)}{\partial b} &= \sum_{n=1}^{N} 2\,(y_i - ax_i - b)^{2-1} \cdot (-1).
\end{aligned}$$

Linear Regression

Explicit formula for values of a, b minimizing error $E_3(a, b)$. From

$$\partial E_3(a,b)/\partial a \;=\; \partial E_3(a,b)/\partial b \;=\; 0 :$$

Expanding gives two linear equations in two unknowns (a, b):

$$\begin{aligned} E_3(a,b) &= \sum_{n=1}^{N} (y_i - ax_i - b)^2 \\ 0 &= \sum_{n=1}^{N} (-x_i y_i) + a \sum_{n=1}^{N} x_i^2 + b \sum_{n=1}^{N} x_i \\ 0 &= \sum_{n=1}^{N} (-y_i) + a \sum_{n=1}^{N} x_i + b \sum_{n=1}^{N} 1. \end{aligned}$$

Linear Regression

Explicit formula for values of a, b minimizing error $E_3(a, b)$. From

$$\partial E_3(a,b)/\partial a \;=\; \partial E_3(a,b)/\partial b \;=\; 0:$$

After algebra:

$$\begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N x_i^2 & \sum_{n=1}^N x_i \\ \sum_{n=1}^N x_i & \sum_{n=1}^N 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=1}^N x_i y_i \\ \sum_{n=1}^N y_i \end{pmatrix}$$

or

$$\begin{aligned} a &= \frac{\sum_{n=1}^N 1 \sum_{n=1}^N x_n y_n - \sum_{n=1}^N x_n \sum_{n=1}^N y_n}{\sum_{n=1}^N 1 \sum_{n=1}^N x_n^2 - \sum_{n=1}^N x_n \sum_{n=1}^N x_n} \\ b &= \frac{\sum_{n=1}^N x_n \sum_{n=1}^N x_n y_n - \sum_{n=1}^N x_n^2 \sum_{n=1}^N y_n}{\sum_{n=1}^N x_n \sum_{n=1}^N x_n - \sum_{n=1}^N x_n^2 \sum_{n=1}^N 1}. \end{aligned}$$

Logarithms and Applications

Many non-linear relationships are linear after applying logarithms:

$$Y = BX^a \text{ then } \log(Y) = a\log(X) + b, \quad b = \log B.$$

Logarithms and Applications

Many non-linear relationships are linear after applying logarithms:

$$Y = BX^a \text{ then } \log(Y) = a\log(X) + b, \quad b = \log B.$$

Kepler's Third Law: if T is the orbital period of a planet traveling in an elliptical orbit about the sun (and no other objects exist), then $T^2 = \widetilde{B}L^3$, where L is the length of the semi-major axis.

Assume do not know this – can we *discover* through statistics?

Logarithms and Applications

Many non-linear relationships are linear after applying logarithms:

$$Y \;=\; BX^a \;\text{ then }\; \log(Y) \;=\; a\log(X)+b, \quad b=\log B.$$

Kepler's Third Law: if T is the orbital period of a planet traveling in an elliptical orbit about the sun (and no other objects exist), then $T = BL^{1.5}$, where L is the length of the semi-major axis.

Assume do not know this – can we *discover* through statistics?

Kepler's Third Law: Can see the 1.5 exponent!

Data: Semi-major axis: Mercury 0.387, Venus 0.723, Earth 1.000, Mars 1.524, Jupiter 5.203, Saturn 9.539, Uranus 19.182, Neptune 30.06 (the units are astronomical units, where one astronomical unit is 1.496 $\cdot 10^8$ km).

Data: orbital periods (in years) are 0.2408467, 0.61519726, 1.0000174, 1.8808476, 11.862615, 29.447498, 84.016846 and 164.79132.

If $T = BL^a$, what should B equal with this data? Units: bruno, millihelen, slug, smoot, ... See https://en.wikipedia.org/wiki/List_of_humorous_units_of_measurement

Kepler's Third Law: Can see the 1.5 exponent!

Data: Semi-major axis: Mercury 0.387, Venus 0.723, Earth 1.000, Mars 1.524, Jupiter 5.203, Saturn 9.539, Uranus 19.182, Neptune 30.06 (the units are astronomical units, where one astronomical unit is $1.496 \cdot 10^8$ km).

Data: orbital periods (in years) are 0.2408467, 0.61519726, 1.0000174, 1.8808476, 11.862615, 29.447498, 84.016846 and 164.79132.

If $T = BL^a$, what should B equal with this data? Units: bruno, millihelen, slug, smoot, ... See https://en.wikipedia.org/wiki/List_of_humorous_units_of_measurement

Kepler's Third Law: Can see the 1.5 exponent!

If try $\log T = a \log L + b$: best fit values $a \approx 1.49986$, $b \approx 0.000148796$:

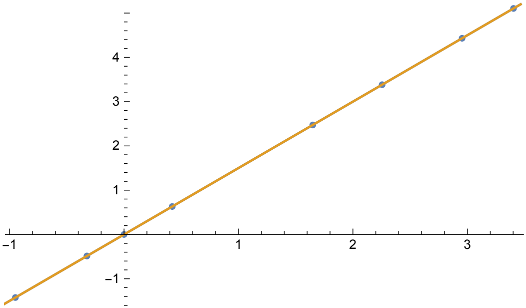


Figure: Plot of $\log P$ versus $\log L$ for planets. Is it surprising $b \approx 0$ (so $B \approx 1$ or $b \approx 0$?

Statistics

Goal is to find good statistics to describe real world.

Figure: Harvard Bridge, about 620.1 meters.

Statistics

Goal is to find good statistics to describe real world.

Figure: Harvard Bridge, 364.1 Smoots ($\pm$ one ear).

Birthday Problem

Birthday Problem: Assume a year with D days, how many people do we need in a room to have a 50% chance that at least two share a birthday, under the assumption that the birthdays are independent and uniformly distributed from 1 to D?

Birthday Problem

Birthday Problem: Assume a year with D days, how many people do we need in a room to have a 50% chance that at least two share a birthday, under the assumption that the birthdays are independent and uniformly distributed from 1 to D?

A straightforward analysis shows the answer is approximately $D^{1/2}\sqrt{\log 4}$.

Can do simulations and try and see the correct exponent; will look not for 50% chance but the expected number of people in room for the first collision.

Birthday Problem (cont)

Try $P = BD^a$, take logs so $\log P = a \log D + b$ ($b = \log B$).

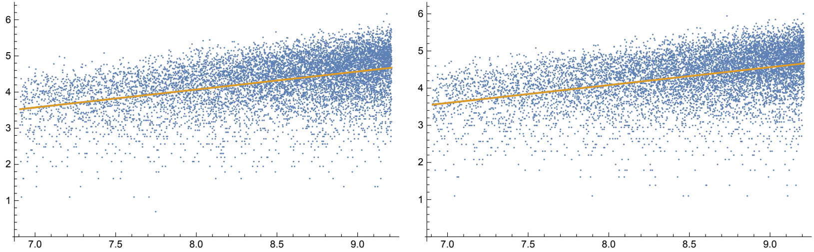


Figure: Plot of best fit line for P as a function of D. We twice ran 10,000 simulations with D chosen from $10,000$ to $100,000$. Best fit values were $a \approx 0.506167$, $b \approx -0.0110081$ (left) and $a \approx 0.48141$, $b \approx 0.230735$ (right).

Statistics and German Tank Problem

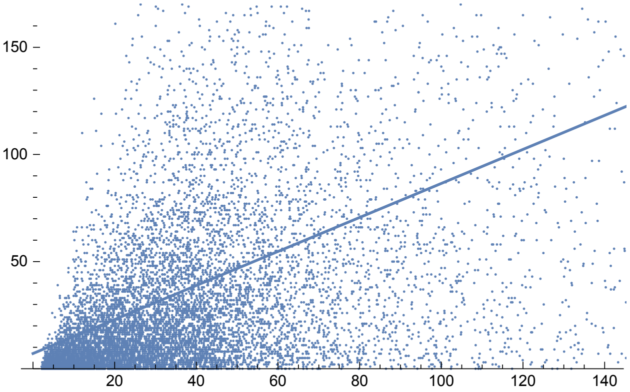


Figure: Plot of best fit line for $N - m$ as a function of m/k. We ran 10,000 simulations with N chosen from $[100, 2000]$ and k from $[10, 50]$. Best fit values for $N - m = a(m/k) + b$ for this simulation were $a \approx 0.793716$, $b \approx 7.10602$.

Statistics and German Tank Problem

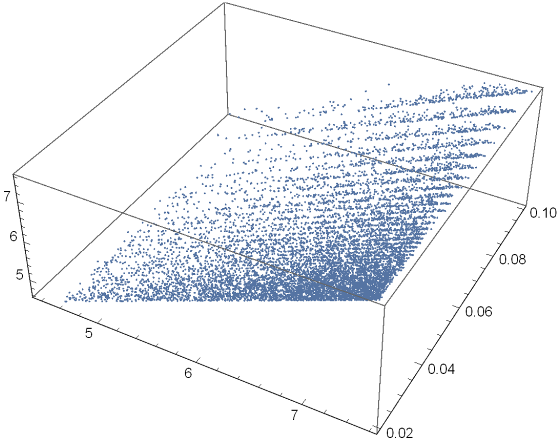


Figure: Plot of $\log N$ against $\log m$ and $1/k$. We ran 10,000 simulations with N chosen from $[100, 2000]$ and k from $[10, 50]$. The data is well-approximated by a plane (we do not draw it in order to prevent our image from being too cluttered).

Implementation Issues

See https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/MethodLeastSquares.pdf

Statistics and German Tank Problem: $N = (1 + 1/k)m - 1$

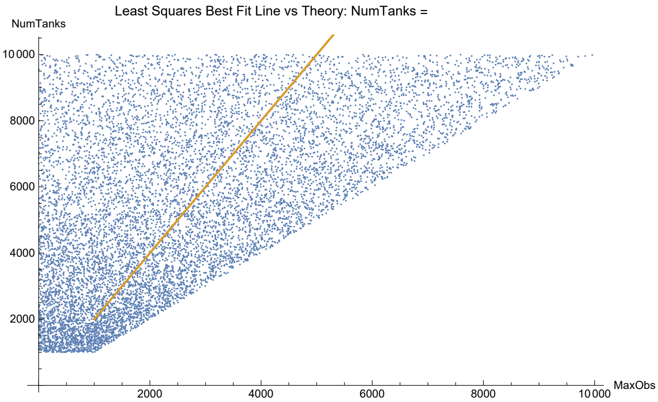


Figure: Plot of N vs maximum observed tank m for fixed $k = 1$. Theory: $N = 2m - 1$, best fit $N = .784m + 2875$.

Statistics and German Tank Problem: $N = (1 + 1/k)m - 1$

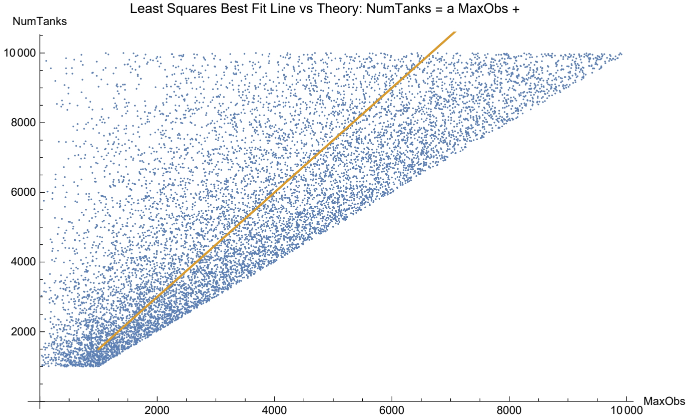


Figure: Plot of N vs maximum observed tank m for fixed $k = 2$. Theory: $N = 1.5m - 1$, best fit $N = .964m + 1424$.

Statistics and German Tank Problem: $N = (1 + 1/k)m - 1$

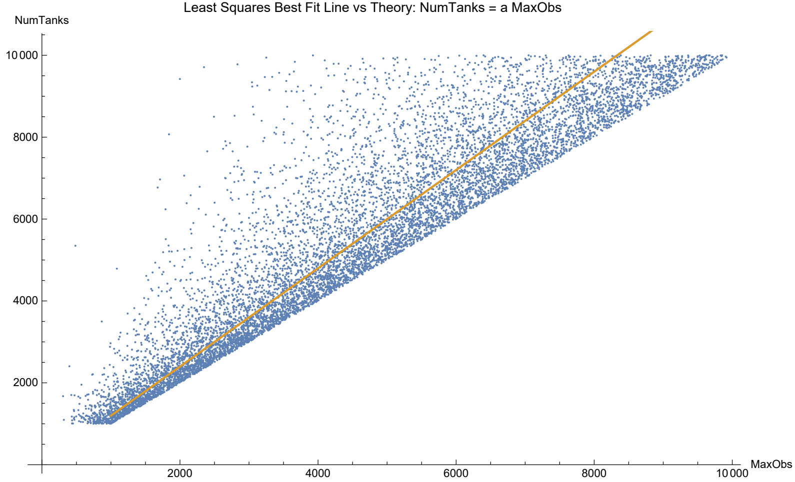


Figure: Plot of N vs maximum observed tank m for fixed $k = 5$. Theory: $N = 1.2m - 1$, best fit $N = 1.037m + 749$.

Statistics and German Tank Problem: $m = \frac{k+1}{k}N + \frac{k+1}{k}$

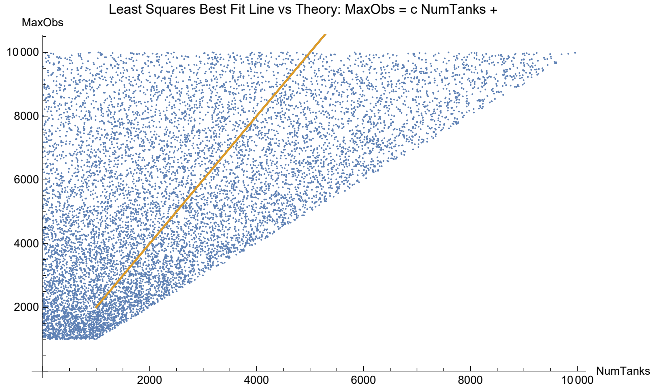


Figure: Plot of maximum observed tank m vs N for fixed $k = 1$. Theory: $m = .5N + .5$, best fit $m = .496N + 10.5$.

Statistics and German Tank Problem: $m = \frac{k+1}{k}N + \frac{k+1}{k}$

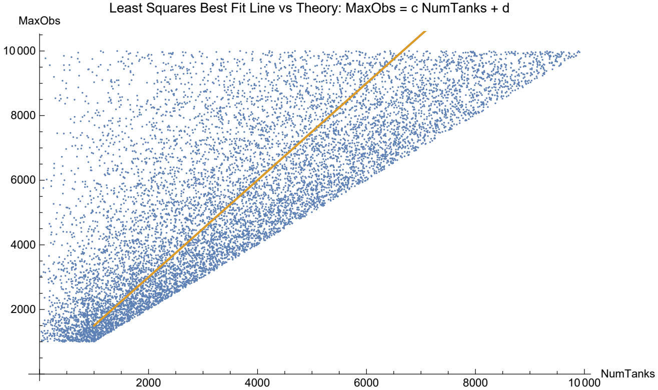


Figure: Plot of maximum observed tank m vs N for fixed $k = 2$. Theory: $m = .667N + .667$, best fit $m = .668N + 9.25$.

Statistics and German Tank Problem: $m = \frac{k+1}{k}N + \frac{k+1}{k}$

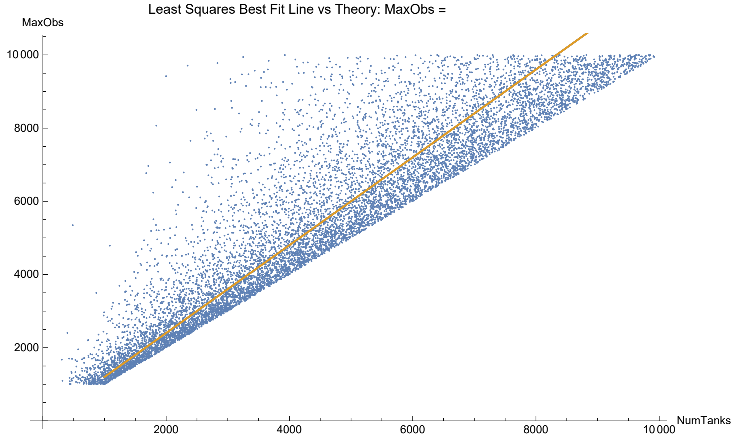


Figure: Plot of maximum observed tank m vs N for fixed $k = 5$. Theory: $m = .883N + .883$, best fit $m = .828N + 25.8$.

Statistics and German Tank Problem: $N = (1 + 1/k)m - 1$: Averaging 100 runs for each N

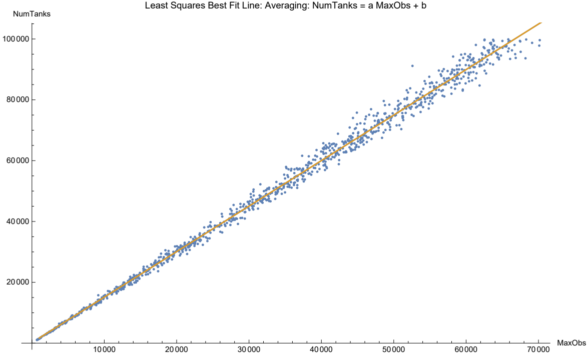


Figure: Plot of N vs maximum observed tank m for fixed $k = 1$. Theory: $N = 1.5m - 1$, best fit $N = 1.496m + 171.2$.

Statistics and German Tank Problem: $N = (1 + 1/k)m - 1$: Sniffing out k dependence

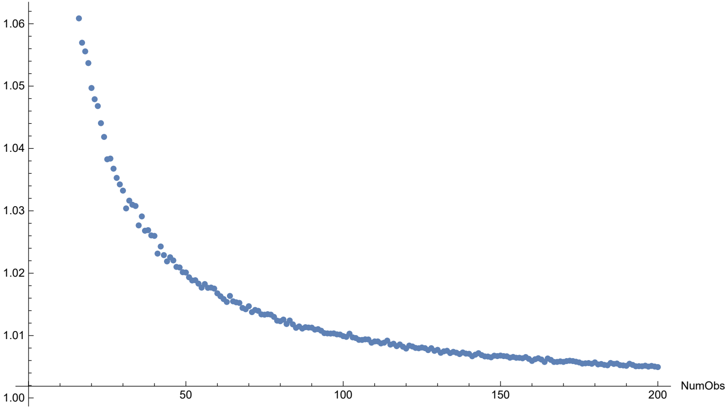


Figure: Plot of a, the slope in $N = am + b$, versus k.

Statistics and German Tank Problem: $N = (1 + 1/k)m - 1$: Sniffing out k dependence

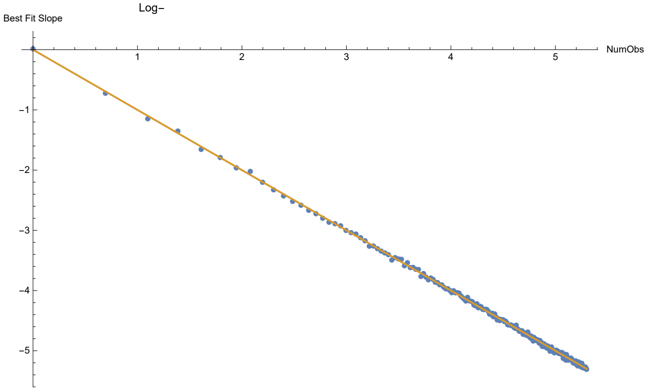


Figure: Log-Log Plot of $a - 1$, the slope in $N = am + b$, versus k. In $\log(a-1)$ versus $\log k$, theory is $\log(a-1) = -1\log(k)$, best fit line is $\log(a-1) = -.999\log(k) - .007$.

References

Smoots

Sieze opportunities: Never know where they will lead.

Smoots

Sieze opportunities: Never know where they will lead.

Smoots

Sieze opportunities: Never know where they will lead.

Oliver Smoot: Chairman of the American National Standards Institute (ANSI) from 2001 to 2002, President of the International Organization for Standardization (ISO) from 2003 to 2004.

Thank you!
Questions?

Work supported by NSF Grants DMS1561945 and DMS1659037.

1. S. J. Miller, *The Probability Lifesaver*, Princeton University Press, Princeton, NJ, 2018. https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/index.htm.
2. Probability and Statistics Blog, *How many tanks? MC testing the GTP*, https://statisticsblog.com/2010/05/25/how-many-tanks-gtp-gets-put-to-the-test/.
3. Statistical Consultants Ltd, *The German Tank Problem*, https://www.statisticalconsultants.co.nz/blog/the-german-tank-problem.html.
4. WikiEducator, *Point Estimation - German Tank Problem*, https://wikieducator.org/Point_estimation_-_German_tank_problem.
5. Wikipedia, *German Tank Problem*, Wikimedia Foundation, https://en.wikipedia.org/wiki/German_tank_problem.

Appendix: Details for Unknown Minimum

Notation

- the minimum tank serial value, N_1,
- the maximum tank serial value, N_2,
- the total number of tanks, N ($N = N_2 - N_1 + 1$),
- the observed minimum value, m_1 (with corresponding random variable M_1),
- the observed maximum value, m_2 (with corresponding random variable M_2),
- the observed spread s (with corresponding random variable S).

As $s = m_2 - m_1$, can focus on just s and S.

Proof

We argue similarly as before. In the algebra below we use our second binomial identity; relabeling the parameters it is

$$\sum_{\ell=a}^{b} \binom{\ell}{a} \;=\; \binom{b+1}{a+1}. \tag{1}$$

We begin by computing the expected value of the spread:

$$\begin{aligned}
\mathbb{E}[S] &= \sum_{s=k-1}^{N-1} s \mathrm{Prob}(S=s) \\
&= \sum_{s=k-1}^{N-1} s \frac{(N-s)\binom{s-1}{k-2}}{\binom{N}{k}} \\
&= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} s\,(N-s)\binom{s-1}{k-2} \\
&= \binom{N}{k}^{-1} N \sum_{s=k-1}^{N-1} \frac{s(s-1)!}{(s-k+1)!(k-2)!} - \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{s^2(s-1)!}{(s-k+1)!(k-2)!} \\
&= \binom{N}{k}^{-1} N \sum_{s=k-1}^{N-1} \frac{s!(k-1)}{(s-k+1)!(k-1)!} - \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{ss!(k-1)}{(s-k+1)!(k-1)!} \;=\; T_1 - T_2.
\end{aligned}$$

Proof (cont)

We first simplify T_1; below we always try to multiply by 1 in such a way that we can combine ratios of factorials into binomial coefficients:

$$\begin{aligned} T_1 &= \binom{N}{k}^{-1} N \sum_{s=k-1}^{N-1} \frac{s!(k-1)}{(s-k+1)!(k-1)!} \\ &= \binom{N}{k}^{-1} N \sum_{s=k-1}^{N-1} \frac{s!(k-1)}{(s-k+1)!(k-1)!} \\ &= \binom{N}{k}^{-1} N(k-1) \sum_{s=k-1}^{N-1} \binom{s}{k-1} \\ &= \binom{N}{k}^{-1} N(k-1)\binom{N}{k} = N(k-1), \end{aligned}$$

where we used (1) with $a = k - 1$ and $b = N - 1$.

Proof (cont)

Turning to T_2 we argue similarly, at one point replacing s with $(s-1)+1$ to assist in collecting factors into a binomial coefficient:

$$
\begin{aligned}
T_2 &= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{s s!(k-1)}{(s-k+1)!(k-1)!} \\
&= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{(s+1-1)s!(k-1)}{(s-(k-1))!(k-1)!} \\
&= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{(s+1)!(k-1)k}{(s+1-k)!(k-1)!k} - \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{s!(k-1)}{(s-(k-1))!(k-1)!} \\
&= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} \frac{(s+1)!k(k-1)}{(s+1-k)!k!} - \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} (k-1)\binom{s}{k-1} \\
&= \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} k(k-1)\binom{s+1}{k} - \binom{N}{k}^{-1} \sum_{s=k-1}^{N-1} (k-1)\binom{s}{k-1} = T_{21} + T_{22}.
\end{aligned}
$$

Proof (cont)

We can immediately evaluate T_{22} by using (1) with $a = k - 1$ and $b = N - 1$, and find

$$T_{22} = \binom{N}{k}^{-1}(k-1)\binom{N}{k} = k - 1.$$

Thus all that remains is analyzing T_{21}:

$$T_{21} = \binom{N}{k}^{-1}\sum_{s=k-1}^{N-1}\binom{s+1}{k}k(k-1).$$

We pull $k(k-1)$ outside the sum, and letting $w = s + 1$ we see that

$$T_{21} = \binom{N}{k}^{-1}k(k-1)\sum_{w=k}^{N}\binom{w}{k},$$

and then from (1) with $a = k$ and $b = N$ we obtain

$$T_{21} = \binom{N}{k}^{-1}k(k-1)\sum_{w=k}^{N}\binom{w}{k} = \binom{N}{k}^{-1}k(k-1)\binom{N+1}{k+1}.$$

Proof (cont)

Thus, substituting everything back yields

$$\mathbb{E}[S] = N(k-1) + (k-1) - \binom{N}{k}^{-1} k(k-1)\binom{N+1}{k+1}.$$

We can simplify the right hand side:

$$\begin{aligned}
(N+1)(k-1) - k(k-1)\frac{\frac{(N+1)!}{(N-k)!(k+1)!}}{\frac{N!}{(N-k)!k!}} &= (N+1)(k-1) - k(k-1)\frac{(N+1)!(N-k)!k!}{N!(N-k)!(k+1)!} \\
&= (N+1)(k-1) - k(k-1)\frac{N+1}{k+1} \\
&= (N+1)(k-1) - \frac{k(k-1)(N+1)}{k+1} \\
&= (N+1)(k-1)\left(1 - \frac{k}{k+1}\right) \\
&= (N+1)\frac{k-1}{k+1},
\end{aligned}$$

and thus obtain

$$\mathbb{E}[S] = (N+1)\frac{k-1}{k+1}.$$

Proof (cont)

The analysis is completed as before, where we pass from our observation of s for S to a prediction $\widehat{N}$ for N:

$$\widehat{N} = \frac{k+1}{k-1}s - 1 = s\left(1 + \frac{2}{k-1}\right) - 1,$$

where the final equality is due to rewriting the algebra to mirror more closely the formula from the case where the first tank is numbered 1. Note that this formula passes the same sanity checks the other did; for example $s\frac{2}{k-1} - 1$ is always at least 1 (remember k is at least 2), and thus the lowest estimate we can get for the number of tanks is $s + 1$.