# The German Tank Problem

# The German Tank Problem

Some statistics for today... :)

# The problem

# The problem

In WW2 allied forces wanted to estimate German tank forces.

# The problem

In WW2 allied forces wanted to estimate German tank forces.
Wikipedia compares for June 1940:

| Statistical | Intelligence | Real |
|---|---|---|
| 159 | 1,000 | 122 |

# Formulation

# Formulation

- Serial numbers on tanks: $1, 2, \ldots, N$.

# Formulation

- Serial numbers on tanks: $1, 2, \ldots, N$.
- Observe: $1 \leq a_1 < a_2 < \ldots < a_k \leq N$.

# Formulation

- Serial numbers on tanks: $1, 2, \ldots, N$.
- Observe: $1 \leq a_1 < a_2 < \ldots < a_k \leq N$.
- **Goal**: Estimate $N$.

# ML-estimator

Let $m = \max_{1 \leq i \leq k} a_i$.

# ML-estimator

Let $m = \max_{1 \leq i \leq k} a_i$.

$$\mathcal{L}(N; m) = \begin{cases} 0, & m > n; \\ \frac{\binom{m-1}{k-1}}{\binom{N}{k}}, & \text{otherwise.} \end{cases}$$

# ML-estimator

Let $m = \max_{1 \leq i \leq k} a_i$.

$$\mathcal{L}(N; m) = \begin{cases} 0, & m > n; \\ \frac{\binom{m-1}{k-1}}{\binom{N}{k}}, & \text{otherwise.} \end{cases}$$

Thus $\hat{N}_{\text{mle}} = m$.

# ML-estimator

Let $m = \max_{1 \leq i \leq k} a_i$.

$$\mathcal{L}(N; m) = \begin{cases} 0, & m > n; \\ \frac{\binom{m-1}{k-1}}{\binom{N}{k}}, & \text{otherwise.} \end{cases}$$

Thus $\hat{N}_{\text{mle}} = m$. Obviously biased: too low for small $k$.

# Frequentionist

$$\hat{N} = \text{maximum observed} + \text{average gap}$$

# Frequentionist

$$\begin{aligned}\hat{N} &= \text{maximum observed} + \text{average gap} \\ &= m + \frac{m}{k} - 1\end{aligned}$$

# Frequentionist

$$\begin{aligned}\hat{N} &= \text{maximum observed} + \text{average gap}\\ &= m + \frac{m}{k} - 1\end{aligned}$$

$\hat{N}$ in between $k$ and $N(1 + \frac{1}{k}) - 1$

# Frequentionist

$$\begin{aligned}\hat{N} &= \text{maximum observed} + \text{average gap} \\ &= m + \frac{m}{k} - 1\end{aligned}$$

$\hat{N}$ in between $k$ and $N(1 + \frac{1}{k}) - 1$

Example: 6 29 8 45. $m = 45$.

# Frequentionist

$$\begin{aligned}\hat{N} &= \text{maximum observed} + \text{average gap} \\ &= m + \frac{m}{k} - 1\end{aligned}$$

$\hat{N}$ in between $k$ and $N(1 + \frac{1}{k}) - 1$

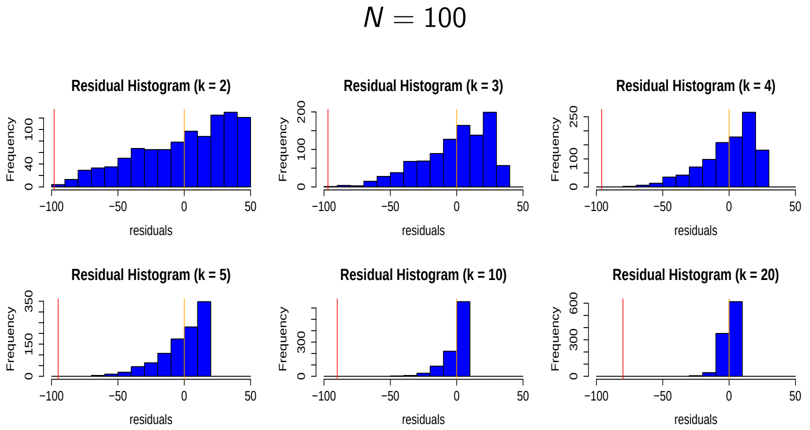
Example: 6 29 8 45. $m = 45$.
$\hat{N} = 45 + 45/4 - 1 = 45 + 11 - 1 = 55$

# Frequentionist

$$\hat{N} = \text{maximum observed} + \text{average gap}$$
$$= m + \frac{m}{k} - 1$$

$\hat{N}$ in between $k$ and $N(1 + \frac{1}{k}) - 1$

Example: 6 29 8 45. $m = 45$.
$\hat{N} = 45 + 45/4 - 1 = 45 + 11 - 1 = 55$ ($N$ was 50).


N = 100
Residual Histogram (k = 2)
Residual Histogram (k = 3)
Residual Histogram (k = 4)
Residual Histogram (k = 5)
Residual Histogram (k = 10)
Residual Histogram (k = 20)
Frequency
residuals
−100
−50
0
50
0
40
100
200
250
150
350
300
600

Thanks!

Thanks!

There is much more to the problem.

Thanks!

There is much more to the problem.
Slides and R code at `github.com/uberwach`