

Same likelihood, different sampling distributions

The MLE of θ under Bernoulli, binomial and negative binomial sampling

Hedibert Freitas Lopes

Inspire

Three sampling rules, one likelihood

- **Bernoulli:** $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Ber}(\theta)$, n fixed

$$L(\theta; y) = \theta^{\sum_i y_i} (1 - \theta)^{n - \sum_i y_i}$$

- **Binomial:** $S \sim \text{Bin}(n, \theta)$, n fixed

$$L(\theta; s) = \binom{n}{s} \theta^s (1 - \theta)^{n-s} \propto \theta^s (1 - \theta)^{n-s}$$

- **Negative binomial:** $N =$ trials until the s -th success, s fixed

$$L(\theta; n) = \binom{n-1}{s-1} \theta^s (1 - \theta)^{n-s} \propto \theta^s (1 - \theta)^{n-s}$$

Three sampling rules, one likelihood

- **Bernoulli:** $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Ber}(\theta)$, n fixed

$$L(\theta; y) = \theta^{\sum_i y_i} (1 - \theta)^{n - \sum_i y_i}$$

- **Binomial:** $S \sim \text{Bin}(n, \theta)$, n fixed

$$L(\theta; s) = \binom{n}{s} \theta^s (1 - \theta)^{n-s} \propto \theta^s (1 - \theta)^{n-s}$$

- **Negative binomial:** $N =$ trials until the s -th success, s fixed

$$L(\theta; n) = \binom{n-1}{s-1} \theta^s (1 - \theta)^{n-s} \propto \theta^s (1 - \theta)^{n-s}$$

Likelihood principle: all three are in the same equivalence class, so any likelihood-based (or Bayesian) inference about θ is the same.

But: the *random* quantity is different (S vs. N), so the sampling distribution of anything computed from L is different.

The MLE is the same function of the data

In all three cases

$$\ell(\theta) = s \log \theta + (n-s) \log(1-\theta), \quad \ell'(\theta) = \frac{s}{\theta} - \frac{n-s}{1-\theta} = 0 \implies \hat{\theta} = \frac{s}{n}.$$

Observed information (identical in all three cases):

$$J(\theta) = -\ell''(\theta) = \frac{s}{\theta^2} + \frac{n-s}{(1-\theta)^2}, \quad J(\hat{\theta}) = \frac{n^3}{s(n-s)} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

The MLE is the same function of the data

In all three cases

$$\ell(\theta) = s \log \theta + (n-s) \log(1-\theta), \quad \ell'(\theta) = \frac{s}{\theta} - \frac{n-s}{1-\theta} = 0 \implies \hat{\theta} = \frac{s}{n}.$$

Observed information (identical in all three cases):

$$J(\theta) = -\ell''(\theta) = \frac{s}{\theta^2} + \frac{n-s}{(1-\theta)^2}, \quad J(\hat{\theta}) = \frac{n^3}{s(n-s)} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

What differs is *which* of (s, n) is random:

	fixed	random
Bern./Binomial	n	$S \Rightarrow \hat{\theta} = S/n$
Neg. binomial	s	$N \Rightarrow \hat{\theta} = s/N$

Bernoulli and binomial: $\hat{\theta} = S/n$

Exact distribution. $S = \sum_i Y_i \sim \text{Bin}(n, \theta)$, hence

$$P\left(\hat{\theta} = \frac{k}{n}\right) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \quad k = 0, 1, \dots, n,$$

with $E(\hat{\theta}) = \theta$ (unbiased) and $\text{Var}(\hat{\theta}) = \theta(1 - \theta)/n$.

Expected (Fisher) information. Using $E(S) = n\theta$:

$$I(\theta) = E\left[\frac{S}{\theta^2} + \frac{n - S}{(1 - \theta)^2}\right] = \frac{n}{\theta} + \frac{n}{1 - \theta} = \frac{n}{\theta(1 - \theta)}.$$

Asymptotics (CLT = MLE theory here):

$$\boxed{\hat{\theta} \stackrel{d}{\sim} N\left(\theta, \frac{\theta(1 - \theta)}{n}\right)}$$

The Bernoulli sample and the binomial count give *identical* answers: S is sufficient, and the order of the y_i 's carries no information.

Negative binomial: $\hat{\theta} = s/N$

Exact distribution. $P(N = m) = \binom{m-1}{s-1} \theta^s (1 - \theta)^{m-s}$, $m = s, s + 1, \dots$,
so

$$P\left(\hat{\theta} = \frac{s}{m}\right) = \binom{m-1}{s-1} \theta^s (1 - \theta)^{m-s}, \quad m = s, s + 1, \dots$$

Support $\{1, s/(s + 1), s/(s + 2), \dots\}$: never 0, accumulates near 0.

Fisher information. Using $E(N) = s/\theta$, $E(N - s) = s(1 - \theta)/\theta$:

$$I(\theta) = \frac{s}{\theta^2} + \frac{s(1 - \theta)/\theta}{(1 - \theta)^2} = \frac{s}{\theta^2(1 - \theta)}.$$

Asymptotics ($s \rightarrow \infty$):

$$\hat{\theta} \stackrel{a}{\sim} N\left(\theta, \frac{\theta^2(1 - \theta)}{s}\right)$$

Same answer via the delta method

N is a sum of s iid $\text{Geometric}(\theta)$ variables, so

$$N \stackrel{a}{\sim} N\left(\frac{s}{\theta}, \frac{s(1-\theta)}{\theta^2}\right).$$

With $g(x) = s/x$, $g'(x) = -s/x^2$ and $g'(s/\theta) = -\theta^2/s$:

$$\text{Var}(\hat{\theta}) \approx \frac{\theta^4}{s^2} \cdot \frac{s(1-\theta)}{\theta^2} = \frac{\theta^2(1-\theta)}{s}. \quad \checkmark$$

Same answer via the delta method

N is a sum of s iid $\text{Geometric}(\theta)$ variables, so

$$N \stackrel{a}{\sim} N\left(\frac{s}{\theta}, \frac{s(1-\theta)}{\theta^2}\right).$$

With $g(x) = s/x$, $g'(x) = -s/x^2$ and $g'(s/\theta) = -\theta^2/s$:

$$\text{Var}(\hat{\theta}) \approx \frac{\theta^4}{s^2} \cdot \frac{s(1-\theta)}{\theta^2} = \frac{\theta^2(1-\theta)}{s}. \quad \checkmark$$

Remark. Plugging in $\hat{\theta} = s/n$:

$$\frac{\hat{\theta}^2(1-\hat{\theta})}{s} = \frac{(s/n)^2(1-\hat{\theta})}{s} = \frac{\hat{\theta}(1-\hat{\theta})}{n}.$$

The *plug-in* standard errors coincide (both equal $J(\hat{\theta})^{-1/2}$), so Wald intervals are identical. The *exact* distributions are not.

Bias: binomial vs. negative binomial

- Binomial: $E(S/n) = \theta$ exactly.
- Negative binomial: by Jensen, $E(s/N) > s/E(N) = \theta$; $\hat{\theta}$ is biased *upward*.
- For $s = 1$ (geometric) there is a closed form:

$$E\left(\frac{1}{N}\right) = \sum_{m \geq 1} \frac{\theta(1-\theta)^{m-1}}{m} = \frac{\theta}{1-\theta} (-\log \theta).$$

At $\theta = 0.1$: $E(\hat{\theta}) \approx 0.256$, more than *twice* θ .

- The unbiased estimator under inverse sampling is $\tilde{\theta} = (s-1)/(N-1)$ (Haldane, 1945).

Bias: binomial vs. negative binomial

- Binomial: $E(S/n) = \theta$ exactly.
- Negative binomial: by Jensen, $E(s/N) > s/E(N) = \theta$; $\hat{\theta}$ is biased *upward*.
- For $s = 1$ (geometric) there is a closed form:

$$E\left(\frac{1}{N}\right) = \sum_{m \geq 1} \frac{\theta(1-\theta)^{m-1}}{m} = \frac{\theta}{1-\theta} (-\log \theta).$$

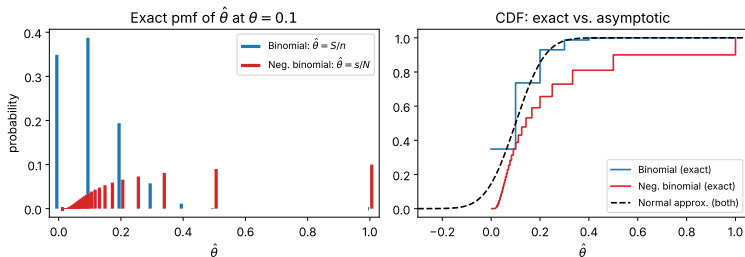
At $\theta = 0.1$: $E(\hat{\theta}) \approx 0.256$, more than *twice* θ .

- The unbiased estimator under inverse sampling is $\tilde{\theta} = (s-1)/(N-1)$ (Haldane, 1945).

Moral: unbiasedness is a property of the sampling rule, not of the likelihood.

Data: $s = 1$ success in $n = 10$ trials

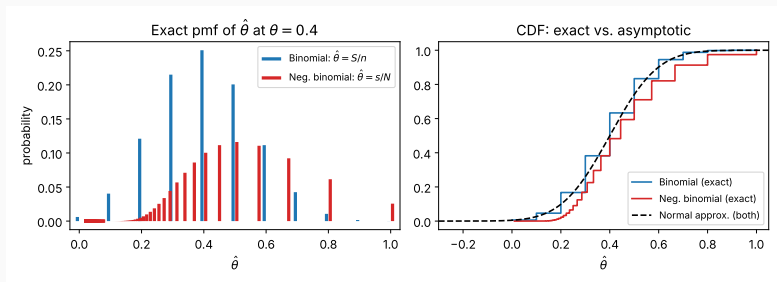
$$\hat{\theta} = 0.1, \quad \widehat{\text{se}} = \sqrt{0.1 \times 0.9/10} = 0.095 \text{ (both rules)}$$



- Binomial: $\hat{\theta} \in \{0, 0.1, \dots, 1\}$; exact sd = 0.095.
- Neg. binomial: $\hat{\theta} \in \{1, \frac{1}{2}, \frac{1}{3}, \dots\}$; exact sd = 0.281, mean 0.256.

Data: $s = 4$ successes in $n = 10$ trials

$$\hat{\theta} = 0.4, \quad \widehat{\text{se}} = \sqrt{0.4 \times 0.6/10} = 0.155 \text{ (both rules)}$$



- Binomial: exact sd = 0.155, unbiased.
- Neg. binomial: exact sd = 0.178, mean 0.461. The normal approximation is already reasonable for the binomial, still skewed for the negative binomial.

Summary of the two data sets (evaluated at $\theta = \hat{\theta}$)

	$s = 1, n = 10$		$s = 4, n = 10$	
	Bin	NegBin	Bin	NegBin
$\hat{\theta}$	0.100	0.100	0.400	0.400
Asymptotic se	0.095	0.095	0.155	0.155
Exact $E(\hat{\theta})$	0.100	0.256	0.400	0.461
Exact $sd(\hat{\theta})$	0.095	0.281	0.155	0.178
Wald 95% CI	$(-0.086, 0.286)$		$(0.096, 0.704)$	

- With $s = 1$ the Wald interval leaves $[0, 1]$: asymptotics are useless here.
- The asymptotic se is blind to the sampling rule; the exact distribution is not.

The classic consequence: p-values

Test $H_0 : \theta = 0.5$ vs. $H_1 : \theta < 0.5$ (Lindley & Phillips, 1976).

- Binomial: “extreme” means *few successes*:
 $p = P(S \leq s \mid n, \theta = 0.5)$.
- Neg. binomial: “extreme” means *many trials*:
 $p = P(N \geq n \mid s, \theta = 0.5)$.

	$s = 1, n = 10$	$s = 4, n = 10$
Binomial p -value	0.0107	0.377
Neg. binomial p -value	0.0020	0.254

Same data, same likelihood, different evidence “against” H_0 , because the reference set of unobserved outcomes depends on the experimenter’s intentions.

Bayesian contrast

With prior $\theta \sim \text{Beta}(a, b)$, under *any* of the three rules

$$\theta \mid \text{data} \sim \text{Beta}(a + s, b + n - s).$$

Uniform prior ($a = b = 1$)	$s = 1, n = 10$	$s = 4, n = 10$
Posterior	$\text{Beta}(2, 10)$	$\text{Beta}(5, 7)$
Posterior mean	0.167	0.417

The stopping rule is irrelevant to the posterior: it enters only through constants $\binom{n}{s}$ or $\binom{n-1}{s-1}$ that cancel.

Take-home: likelihood-based conclusions obey the likelihood principle; frequentist properties of $\hat{\theta}$ (bias, exact variance, p -values, coverage) do not.