Question 1

Naman Agrawal

2025-01-03

The geometric distribution is ideal for modeling the expected number of tries before successfully grabbing a tub of ice cream because it describes the probability of observing the first success in a series of Bernoulli trials. In this scenario:

1. Each trial (playing the machine) is independent.
2. Each trial has two outcomes: success (winning a tub) or failure (not winning a tub).
3. The probability of success in each trial is constant.

If p is the probability of success in one trial, the expected number of tries X is modeled as:

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

The expected value is:

$$E[X] = \frac{1}{p}.$$

Estimation of p

To compute p from actual data, we can use three approaches: Maximum Likelihood Estimation (MLE), Method of Moments (which coincides with MLE here), and Bayesian Estimation.

Maximum Likelihood Estimation (MLE)

The likelihood function is:

$$L(p; X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} (1 - p)^{X_i - 1} p$$

The log-likelihood is:

$$\ell(p; X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} [(X_i - 1) \ln(1 - p) + \ln(p)]$$

Differentiating and solving for p:

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Method of Moments

Equating the sample mean $\bar{X}$ to the theoretical mean $\frac{1}{p}$:

$$\hat{p} = \frac{1}{\bar{X}} = \frac{n}{\sum_{i=1}^n X_i}$$

Bayesian Estimation

With a Beta prior:

$$p \sim \text{Beta}(\alpha, \beta),$$

The prosterior is given by:

$$\pi(p \mid X_1, X_2, \ldots, X_n) \propto \pi(X_1, X_2, \ldots, X_n \mid p) \cdot \pi(p)$$

$$\propto \prod_{i=1}^{n} \left[(1-p)^{X_i - 1} p\right] \cdot p^{\alpha-1}(1-p)^{\beta-1}$$

$$\propto (1-p)^{\sum_{i=1}^n X_i - n + \beta - 1} p^{\alpha+n-1}$$

and posterior:

$$p \mid X_1, X_2, \ldots, X_n \sim \text{Beta}\left(\alpha + n, \beta + \sum_{i=1}^{n} X_i - n\right)$$

The posterior mean is:

$$\hat{p}\text{Bayes} = \frac{\alpha + n}{\alpha + \beta + \sum_{i=1}^n X_i}$$

Thus, MLE and Method of Moments provide identical estimates, while Bayesian estimation incorporates prior beliefs for a more robust estimate when prior information is available.

Confidence Intervals (CI)

Confidence Interval for MLE

For the MLE estimator $\hat{p} = \frac{n}{\sum_{i=1}^n X_i}$, the asymptotic variance of the MLE is the inverse of the Fisher Information:

$$I(p) = -\mathbb{E}\left[\frac{\partial^2 \ell(p)}{\partial p^2}\right] = \frac{n}{p^2(1-p)}$$

$$\text{Var}(\hat{p}) = \frac{p^2(1-p)}{n}$$

where n is the number of observations. The standard error is:

$$\text{SE} = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{\hat{p}^2(1-\hat{p})}{n}}$$

A 95% confidence interval is:

$$\hat{p} \pm z_{\alpha/2} \cdot \text{SE},$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution (e.g., $z_{0.025} = 1.96$ for 95% CI).

Since the Method of Moments (MoM) estimator coincides with the MLE, the CI for MoM is identical to that of MLE:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}^2(1-\hat{p})}{n}}$$

Confidence Interval for Bayesian Estimation

Using the posterior distribution $p \mid X_1, X_2, \ldots, X_n \sim \text{Beta}(\alpha + n, \beta + \sum_{i=1}^n X_i - n)$, credible intervals can be computed. The interval $[p_{\text{lower}}, p_{\text{upper}}]$ is chosen such that:

$$P(p_{\text{lower}} \leq p \leq p_{\text{upper}} \mid \text{data}) = 0.95$$

This can be done by finding the 2.5th and 97.5th percentiles of the posterior Beta distribution:

$$p_{\text{lower}} = F^{-1}\text{Beta}(0.025; \alpha + n, \beta + \sum_{i=1}^{n} X_i - n)$$

$$p_{\text{upper}} = F^{-1}\text{Beta}(0.975; \alpha + n, \beta + \sum_{i=1}^{n} X_i - n)$$

Here, F^{-1}_{Beta} is the inverse cumulative distribution function (CDF) of the Beta distribution.

Simulating Data and Computing Estimates with Confidence Intervals in R

Below is an example in R to simulate data for the number of trials before a success and compute estimates for p using the above methods. Confidence intervals for all methods are calculated.

```r
library(MASS)

# simulate data
set.seed(123)
true_p <- 0.2
n <- 100
simulated_data <- rgeom(n, prob = true_p) + 1

cat("simulated data summary:\n")
```

```
## simulated data summary:
```

```r
summary(simulated_data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    4.00    4.78    6.25   16.00
```

```r
cat("true prob of success (p):", true_p, "\n")
```

```
## true prob of success (p): 0.2
```

```r
# MLE and Method of Moments
sample_mean <- mean(simulated_data)
mle_p <- 1 / sample_mean
mom_p <- mle_p
std_error <- mle_p*sqrt((1 - mle_p)/n)
z <- 1.96
ci_mle <- c(mle_p - z * std_error, mle_p + z * std_error)

# Bayesian estimation (uniform prior)
alpha <- 1
beta <- 1
posterior_alpha <- alpha + n
posterior_beta <- beta + sum(simulated_data) - n
bayesian_p <- posterior_alpha / (posterior_alpha + posterior_beta)
ci_bayes <- qbeta(c(0.025, 0.975), posterior_alpha, posterior_beta)

# results
cat("\nComparison of Estimates and CIs:\n")
```

```
##
## Comparison of Estimates and CIs:
```

```r
cat("True p:", true_p, "\n")
```

```
## True p: 0.2
```

```r
cat("MLE p:", mle_p, "CI:", ci_mle, "\n")
```

```
## MLE p: 0.209205 CI: 0.1727414 0.2456687
```

```r
cat("MoM p:", mom_p, "CI:", ci_mle, "\n")
```

```
## MoM p: 0.209205 CI: 0.1727414 0.2456687
```

```r
cat("Bayesian p:", bayesian_p, "CI:", ci_bayes, "\n")
```

```
## Bayesian p: 0.2104167 CI: 0.1751624 0.2479561
```