Normal Approximation

MATH/STAT 394: Probability I
Summer 2021 A Term

Introduction to Probability
D. Anderson, T.Seppäläinen, B. Valkó

§ 4.3

Aaron Osgood-Zimmerman

Department of Statistics

Practice solution

Practice

We roll a pair of fair dice 10,000 times.

Estimate the prob. that the number of times we get snake eye s (two ones) is between 280 and 300

Solution

- Denote X the nb of snake eyes we get in the 10,000 rolls, i.e. $X \sim \mathrm{Bin}(10,000, 1/36)$
- Using the normal approx. with $n = 10,000$, $p = 1/36$,

$$\begin{aligned}\mathbb{P}(280 \leq X \leq 300) &= \mathbb{P}\left(\frac{280-np}{\sqrt{np(1-p)}} \leq \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{300-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{300-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{280-np}{\sqrt{np(1-p)}}\right) \\ &\approx 0.3578\end{aligned}$$

Recap

Standard Normal/Gaussian Distribution

- $Z \sim \mathcal{N}(0,1)$ has p.d.f.
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$
- c.d.f. $\Phi(x)$ not avail. in closed form but given by tables
- $\mathbb{E}[Z] = 0$, $\mathrm{Var}(Z) = 1$

Normal/Gaussian distribution

- $X \sim \mathcal{N}(\mu, \sigma^2)$ has p.d.f.
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$
- $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$

From standard to not-standard and vice-versa

- if $Z \sim \mathcal{N}(0,1)$, then $X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$
- if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$
- More generally if $X \sim \mathcal{N}(\mu, \sigma^2)$, $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

Recap

Normal approximation to Binomial

- If $S_n \sim \text{Bin}(n, p)$, consider the *standardization* of S_n, i.e.,

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

- Then the standardized S_n tends to be a standard normal dist. as $n \to +\infty$, i.e.

$$\lim_{n \to +\infty} \mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

  the fact that S_n converges to a Gaussian is called the **central limit theorem**

- Practically for $np(1-p) > 10$ (i.e. n large, p not too close to 0 or 1)

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

Outline

CLT

We saw the CLT for a Binomial r.v., but it is much more general:

Theorem (Central limit theorem)
Let $X_1, X_2, \ldots, X_n$ be a sequence of n i.i.d r.v.s with mean $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

Let $\bar{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n}$. Then, as n approaches infinity, the random variable $\sqrt{n}(S_n - \mu)$ converges in distribution to a $\mathcal{N}(0, \sigma^2)$. So,n

$$\lim_{n \to +\infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \underset{d}{\to} \mathcal{N}(0, 1)$$

Thus for any $-\infty \leq a \leq b \leq +\infty$,

$$\lim_{n \to +\infty} \mathbb{P}\left(a \leq \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \leq b \right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Interpretation: The average of a large number of r.v.s will tend towards a Gaussian distribution!

Outline

The Central Limit Theorem

Applications of the normal approximation

Confidence intervals

Motivation

- Suppose we have a biased coin and we want to know $p = \mathbb{P}(\text{getting a tail})$
- How can we know if our observed frequency of tails $\hat{p} = \frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n}$ is a good estimate of p?
- We want to estimate for some $\varepsilon > 0$ fixed, the prob. $\mathbb{P}(|p - \hat{p}| \leq \varepsilon)$

Confidence Intervals

- Let us reformulate $\mathbb{P}(|p-\hat{p}| \leq \varepsilon)$ in terms of the central limit theorem

$$\begin{aligned}
\mathbb{P}(|p-\hat{p}| < \varepsilon) &= \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) \\
&= \mathbb{P}(-n\varepsilon < S_n - np < n\varepsilon) \\
&= \mathbb{P}\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \\
&\approx \Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \\
&= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1
\end{aligned}$$

- **Problem:**
  we do not know p so we cannot compute the right hand side...
- **Solution:**
  Upper bound $p(1-p)$, (here $p(1-p) \leq 1/4$ for all $p \in (0,1)$) ,
  then as Φ is increasing, $\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \geq \Phi\left(2\varepsilon\sqrt{n}\right)$ and so

$$\boxed{\mathbb{P}(|p-\hat{p}| < \varepsilon) \geq 2\Phi\left(2\varepsilon\sqrt{n}\right) - 1}$$

Confidence Intervals

Exercise

How many times should you flip a coin with unknown prob. of success p such that the estimate $\hat{p} = \frac{S_n}{n}$ is within 0.05 of the true p with prob. at least 0.99?

Solution

- We need n such that

$$\mathbb{P}(|p - \hat{p}| < \varepsilon) \geq 2\Phi\left(2\varepsilon\sqrt{n}\right) - 1 \geq 0.99$$

  which is satisfied for $\Phi\left(2\varepsilon\sqrt{n}\right) \geq 0.995$

- From a table of the c.d.f. Φ we find that it is equivalent to

$$2\varepsilon\sqrt{n} \geq 2.58 \quad \text{i.e.} \quad n \geq \frac{2.58^2}{4\varepsilon^2} = \frac{2.58^2}{4 \cdot 0.05^2} \approx 665.64$$

- So we need approx. 666 flips to get an estimate $\hat{p}$ within 0.05 of the true p with prob. at least 0.99

Confidence intervals

Definition

Let $X \sim \mathrm{Ber}(p)$ and $\hat{p}$ be an estimator of p.

A confidence interval at level α of p is an interval of the form

$$[\hat{p} - \varepsilon, \hat{p} + \varepsilon] \quad \textit{s.t.} \quad \mathbb{P}(p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]) \geq \alpha$$

which is equivalent to $\mathbb{P}(|p - \hat{p}| \leq \varepsilon) \geq \alpha$.

Note:

- Here the randomness lies in $\hat{p}$ not in p, ε or α
- Such confidence intervals can be computed using the central limit theorem

Confidence Intervals

Exercise

We repeat a trial 1000 times and observe 450 successes.

Find a 95% confidence interval for the true success prob. p

Solution

- Form the previous slides we know that
$$\mathbb{P}(|p - \hat{p}| < \varepsilon) \geq 2\Phi\left(2\varepsilon\sqrt{n}\right) - 1$$
- So we need to find ε s.t. $2\Phi(2\varepsilon\sqrt{n}) - 1 \geq 0.95$ where $n = 1000$ which is equivalent to find ε s.t. $\Phi(2\varepsilon\sqrt{n}) \geq 0.975$
- By looking at a table of Φ we get that this inequality is satisfied for
$$2\varepsilon\sqrt{n} \geq 1.96 \Leftrightarrow \varepsilon \geq \frac{1.96}{2\sqrt{1000}} \approx 0.0031$$
- Therefore plugging $\varepsilon = 0.0031$, and $\hat{p} = 450/1000 = 0.45$ we get that with prob. greater than 0.95
$$|p - 0.45| < 0.031$$
- Namely with prob. greater than 0.95, the true success prob p lies in
$$[0.45 - 0.031, 0.45 + 0.031] = [0.419, 0.481]$$

Practice next lecture

Practice

Suppose we interviewed 400 people and 100 of them liked spinach

Find a 90% confidence interval for the true probability that people like spinach assuming that we may call the same person twice (sampling with replacement)