Welcome to CS109A

Gili Rusak

Agenda

- Central limit theorem*
- Bootstrap method*
- Applications

* Relevant for HW5

Central Limit Theorem

Definition

The Central Limit Theorem (CLT) proves that the averages of samples from *any* distribution themselves must be normally distributed. Consider IID random variables $X_1, X_2 \ldots$ such that $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let

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

The Central Limit Theorem states:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \text{as } n \to \infty$$

Example Problem

You will roll a 6 sided dice 10 times. Let X be the total value of all 10 dice = $X1$ + $X2$ + · · · + $X10$. You win the game if $X \leq 25$ or $X \geq 45$. Use the Central Limit Theorem to calculate the probability that you win.

Example Problem Solution

Step 1:

Recall that $E[X_i] = 3.5$ and $\text{Var}(X_i) = \frac{35}{12}$.

Example Problem Solution

Step 1:

Recall that $E[X_i] = 3.5$ and $\text{Var}(X_i) = \frac{35}{12}$.

Step 2:

$$
\begin{aligned}
P(X \le 25 \text{ or } X \ge 45) &= 1 - P(25.5 \le X \le 44.5) \\
&= 1 - P\left(\frac{25.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \le \frac{X - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \le \frac{44.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}}\right) \\
&\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784
\end{aligned}
$$

Bootstrap Method

Bootstrap Method

```
def bootstrap(sample):
    n = number of elements in sample
    pmf = estimate the underlying pmf from the sample
    stats = []
    repeat 10,000 times:
        resample = draw n new samples from the pmf
        stat = calculate your stat on the resample
        stats.append(stat)
    stats can now be used to estimate the distribution of the stat
```

Example Problem

A medical researcher treats patients with dangerously low hemoglobin levels. She has formulated two slightly different drugs and is now testing them on patients. First, she administered drug A to one group of 50 patients and drug B to a separate group of 50 patients. Then, she measured all the patients' hemoglobin levels post-treatment. For simplicity, assume that all variation in the patient outcomes is due to their different reactions to treatment.

The researcher notes that the sample mean is similar between the two groups: both have mean hemoglobin levels around 10g/dL. However, drug B's group has a **sample variance** that is 3 $(\text{g/dL})^2$ **greater** than drug A's group. The researcher thinks that patients respond to drugs A and B differently. Specifically, she wants to make the scientific claim that drug A's patients will end up with a significantly different spread of hemoglobin levels compared to drug B's.

You are skeptical. It is possible that the two drugs have practically identical effects and that the observed different in variance was a result of chance and a small sample size, i.e. the **null hypothesis**. Calculate the probability of the null hypothesis using bootstrapping. Here is the data. Each number is the level of an independently sampled patient:

Try it out!

https://colab.research.google.com/drive/1kLokD2FPkQo4cpyO2fc_XBDrb7AImMip?usp=sharing

See you next Tuesday!