6.2: The Distribution of the Sample Mean

Taylor

University of Virginia

Proposition

Moments of an i.i.d. Sample

Let $X_1, \ldots, X_n$ be a random sample from any distribution that has a mean and variance. Then:

- $E[\bar{X}] = E[X_i]$
- $E[\sum X_i] = nE[X_i]$
- $V[\bar{X}] = \frac{V[X_i]}{n}$
- $V[\sum X_i] = nV[X_i]$

Proposition

Distribution of an i.i.d. Normal Sample

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Normal}(\mu, \sigma^2)$. Then:

$$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$$

and

$$\sum X_i \sim \text{Normal}(n\mu, n\sigma^2)$$

Example

Example from page 298:

Let $X_1 \ldots X_5$ denote a random sample of rat exit times from a maze (in minutes). Each X_i is distributed as a Normal$(1.5, .35^2)$ rv. If rat's go one after another, what is the probability that the total time is between 6 and 8 minutes?

(I'm not writing out the soluton here. But you use the previous theorems. This question is asking us about $\sum_i X_i$, so we get the distribution of that.)

Note

- we can't always get by assuming our data is normally distributed
- fortunately we have some big theorems in statistics that let us use the tools we're learning about in more general circumstances
- Before I introduce these theorems, we have to talk about some different types of convergence

Three types of convergence

Convergence in Distribution

Y_n converges in distribution to Y (written as $Y_n \xrightarrow{D} Y$) if $P(Y_n \leq c) \to P(Y \leq c)$ for any point of continuity c as $n \to \infty$

Convergence in Mean Square

$\bar{X}_n$ converges in m.s. to θ (written as $\bar{X}_n \xrightarrow{m.s.} \theta$) if $E[(\bar{X}_n - \theta)^2] \to 0$ as $n \to \infty$

Convergence in Probability

$\bar{X}_n$ converges in probability to θ (written as $\bar{X}_n \xrightarrow{p} \theta$) if for any ϵ, $P(|\bar{X}_n - \theta| > \epsilon) \to 0$ as $n \to \infty$.

Theorem

Now we can answer this question: what if we cannot assume the data are normally distributed?

Theorem

Now we can answer this question: what if we cannot assume the data are normally distributed?

a Central Limit Theorem (C.L.T)

Let $X_1, \ldots, X_n$ be a random sample from some distribution that has a mean and variance. Then

$$\frac{\bar{X} - E(X_i)}{\sqrt{\frac{V(X_i)}{n}}} \xrightarrow{D} Z$$

as $n \to \infty$.

Note

A few things:

- In practice this means that if we have a lot of data, we can do the same thing we did before and we don't have to worry about if the data are individually coming from a Normal distribution.
- This is a *very* important theorem
- I'm not going to make up an arbitrary rule of thumb for when n is big enough and have you memorize it

Example

Recall your STAT 2120 hypothesis test for testing a population proportion. Say you take a sample of 100 people and ask them if they are Republican or not. A 'yes' will be denoted by a 1, ($X_i = 1$) and a 'no' will be a 0 ($X_i = 0$). If we assume that this is all an i.i.d. random sample from a Bernoulli(p) distribution, then

$$\sum X_i \sim \text{Binomial}(100, p)$$

...but what about $\bar{X} = \frac{1}{n}\sum X_i$?

Example (continued)

If we assume $p = .5$, then what is the probability that our sample proportion of republicans is larger than .6? n seems big enough here (100), so let's use our CLT. Recall that the mean of a Bernoulli random variable is p and the variance of a Bernoulli rv is $p(1-p)$.

$$
\begin{aligned}
P(\bar{X} > .6) &= P\left(\frac{\bar{X} - .5}{\sqrt{\frac{.25}{100}}} > \frac{.6 - .5}{\sqrt{\frac{.25}{100}}}\right) \\
&\to P\left(Z > \frac{.6 - .5}{\sqrt{\frac{.25}{100}}}\right) \\
&= P(Z > 2) \\
&= 1 - P(Z \leq 2) \\
&= 0.02275013
\end{aligned}
$$

Two Laws of Large Numbers

LLN 1

Let $X_1, \ldots, X_n$ be an i.i.d. sample from some distribution with a mean and variance. Then

$$\bar{X} \xrightarrow{p} E(X_i)$$

LLN 2

Let $X_1, \ldots, X_n$ be an i.i.d. sample from some distribution with a mean and variance. Then

$$\bar{X} \xrightarrow{m.s.} E(X_i)$$

Example

You know a lot about football and you can usually pick which team wins 53% of the time. If you are right, your bookie gives you a thousand dollars. If you are wrong, you pay your bookie a thousand dollars. However, you also have to pay $10 every time you play this game. What happens to your average winnings as you play this game over and over again (and assuming the outcomes are i.i.d.)?

Example

Let $X_1, \ldots X_n$ denote your payouts for your games.

$$E(X_i) = .53 * 1000 + .47 * (-1000) - 10 = 50$$

By either of the LLNs, $\bar{X} \to E(X_i) = 50$. So in the long run you make \$50 a game.

Example

The Monte-Carlo Method

Let $X_1, \ldots, X_n$ be an iid sample from some distribution. Then:

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \xrightarrow{p} E[g(X)].$$

This is just convergence in probability/mean (define $Y_i = g(X_i)$ and apply the theorem to the Ys!)

MC Example

Let $X \sim \text{Normal}(1, 40)$. Say you want to estimate $E[\sin(X)]$.

$$\frac{\sum_{i=1}^{n} \sin(X_i)}{n} \xrightarrow{p} E[\sin(X)]$$

In practice this means simulate $X_1, \ldots, X_{10000}$ from this Normal distribution, and then calculate $\frac{\sum_{i=1}^{n} \sin(X_i)}{n}$.

MC Example

Say you have a dataset $X_1, \ldots, X_n$, a random sample. You don't know what distribution it comes from, but you are only interested in $P(X > 4)$. Then

$$\frac{\sum_{i=1}^{n} 1(X_i > 4)}{n} \xrightarrow{p} E[1(X > 4)] = P(X > 4).$$

In practice this means calculate $\frac{\sum_{i=1}^{n} 1(X_i > 4)}{n}$ by just counting how many values exceed 4, and then divide this number by the total data size.