THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 321: Introduction to Probability and Statistics

Lecture 18: Central limit theorem

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

www.cs.cmu.edu/∼psarkar/teaching

Some useful inequalities

So far we have looked at expectations and variances of sums of independent random variables. Today we will also look at there behavior when the number of random variables is increasing.

- Remember markov's inequality? For a positive random variable X and some $t > 0$, we said that $P(X \geq t) \leq \frac{E[X]}{t}$
- We can use this to bound $P(|X - E[X]| \geq c)$.

$$P(|X - \mu| \geq c) = P((X - \mu)^2 \geq c^2) \leq \frac{E[(X - \mu)^2]}{c^2}$$

- This is the famous Chebyshev inequality.
- All this comes in handy to show that a random variable cannot be too far from its expectation if the variance is small.

Weak law of large numbers

The WLLN basically states that the sample mean of a large number of random variables is very close to the true mean with high probability.

- Consider a sequence of i.i.d random variables $X_1, \ldots X_n$ with mean μ and variance σ^2.
- Let $M_n = \dfrac{X_1 + \cdots + X_n}{n}$.

Weak law of large numbers

The WLLN basically states that the sample mean of a large number of random variables is very close to the true mean with high probability.

- Consider a sequence of i.i.d random variables $X_1, \dots X_n$ with mean μ and variance σ^2.
- Let $M_n = \dfrac{X_1 + \cdots + X_n}{n}$.
- $E[M_n] = \dfrac{E[X_1] + \cdots + E[X_n]}{n} = \mu$

Weak law of large numbers

The WLLN basically states that the sample mean of a large number of random variables is very close to the true mean with high probability.

- Consider a sequence of i.i.d random variables $X_1, \dots X_n$ with mean μ and variance σ^2.
- Let $M_n = \dfrac{X_1 + \cdots + X_n}{n}$.
- $E[M_n] = \dfrac{E[X_1] + \cdots + E[X_n]}{n} = \mu$
- $\text{var}(M_n) = \dfrac{\text{var}[X_1] + \cdots + \text{var}[X_n]}{n^2} = \dfrac{\sigma^2}{n}$

Weak law of large numbers

The WLLN basically states that the sample mean of a large number of random variables is very close to the true mean with high probability.

- Consider a sequence of i.i.d random variables $X_1, \ldots X_n$ with mean μ and variance σ^2.
- Let $M_n = \dfrac{X_1 + \cdots + X_n}{n}$.
- $E[M_n] = \dfrac{E[X_1] + \cdots + E[X_n]}{n} = \mu$
- $\text{var}(M_n) = \dfrac{\text{var}[X_1] + \cdots + \text{var}[X_n]}{n^2} = \dfrac{\sigma^2}{n}$
- So $P(|M_n - \mu| \geq \epsilon) \leq \dfrac{\sigma^2}{n\epsilon^2}$

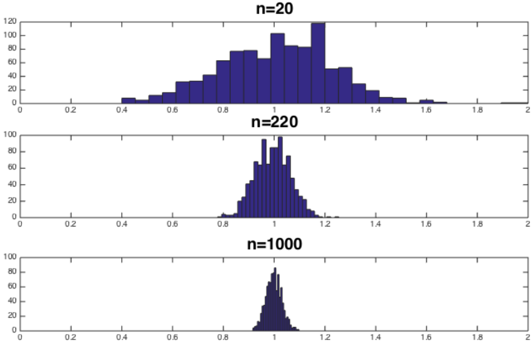
Weak law of large numbers

The WLLN basically states that the sample mean of a large number of random variables is very close to the true mean with high probability.

- Consider a sequence of i.i.d random variables $X_1, \dots X_n$ with mean μ and variance σ^2.
- Let $M_n = \dfrac{X_1 + \cdots + X_n}{n}$.
- $E[M_n] = \dfrac{E[X_1] + \cdots + E[X_n]}{n} = \mu$
- $\text{var}(M_n) = \dfrac{\text{var}[X_1] + \cdots + \text{var}[X_n]}{n^2} = \dfrac{\sigma^2}{n}$
- So $P(|M_n - \mu| \geq \epsilon) \leq \dfrac{\sigma^2}{n\epsilon^2}$
- For large n this probability is small.

Illustration

Consider the mean of n independent Poisson(λ) random variables. For each n, we plot the distribution of the average.



n=20

n=220

n=1000

Can we say more?

Turns out that not only can you say that the sample mean is close to the true mean, you can actually predict its distribution using the famous Central Limit Theorem.

- Consider a sequence of i.i.d random variables $X_1, \ldots X_n$ with mean μ and variance σ^2.
- Let $M_n = \dfrac{X_1 + \cdots + X_n}{n}$. Remember $E[M_n] = \mu$ and var$(M_n) = \sigma^2/n$
- Standardize M_n to get $\dfrac{M_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\dfrac{M_n - \mu}{\sigma}$
- As n gets bigger, $\sqrt{n}\dfrac{M_n - \mu}{\sigma}$ behaves more and more like a $Normal(0,1)$ random variable.
- $P(\sqrt{n}\dfrac{M_n - \mu}{\sigma} < z) \approx \Phi(z)$

Practice problem

Let X_1, Y_1, X_2, Y_2 be independent random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$W = \frac{X_1 + \cdots + X_{16} - (Y_1 + \cdots + Y_{16})}{16}$$

Find a numerical approximation of $P(|W - E[W]| > .001)$.

Practice problem

Let X_1, Y_1, X_2, Y_2 be independent random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$W = \frac{X_1 + \cdots + X_{16} - (Y_1 + \cdots + Y_{16})}{16}$$

Find a numerical approximation of $P(|W - E[W]| > .001)$.

- We can rewrite W as $W = \frac{(X_1 - Y_1) + (X_2 - Y_2) + \cdots + (X_{16} - Y_{16})}{16}$

Practice problem

Let X_1, Y_1, X_2, Y_2 be independent random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$W = \frac{X_1 + \cdots + X_{16} - (Y_1 + \cdots + Y_{16})}{16}$$

Find a numerical approximation of $P(|W - E[W]| > .001)$.

- We can rewrite W as $W = \dfrac{(X_1 - Y_1) + (X_2 - Y_2) + \cdots + (X_{16} - Y_{16})}{16}$
- Now note that W is an average of 16 i.i.d random variables. Let $Z_i = X_i - Y_i$.

Practice problem

Let X_1, Y_1, X_2, Y_2 be independent random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$W = \frac{X_1 + \cdots + X_{16} - (Y_1 + \cdots + Y_{16})}{16}$$

Find a numerical approximation of $P(|W - E[W]| > .001)$.

- We can rewrite W as $W = \frac{(X_1 - Y_1) + (X_2 - Y_2) + \cdots + (X_{16} - Y_{16})}{16}$
- Now note that W is an average of 16 i.i.d random variables. Let $Z_i = X_i - Y_i$.
- $E[Z_i] = E[X_i] - E[Y_i] = 0$ and so $E[W] = 0$.

Practice problem

Let X_1, Y_1, X_2, Y_2 be independent random variables, uniformly distributed in the unit interval $[0, 1]$. Let

$$W = \frac{X_1 + \cdots + X_{16} - (Y_1 + \cdots + Y_{16})}{16}$$

Find a numerical approximation of $P(|W - E[W]| > .001)$.

- We can rewrite W as $W = \frac{(X_1 - Y_1) + (X_2 - Y_2) + \cdots + (X_{16} - Y_{16})}{16}$
- Now note that W is an average of 16 i.i.d random variables. Let $Z_i = X_i - Y_i$.
- $E[Z_i] = E[X_i] - E[Y_i] = 0$ and so $E[W] = 0$.
- $\sigma^2 = \text{var}(Z_i) = \text{var}(X_i - Y_i) = \text{var}(X_i) + \text{var}(Y_i) = 1/12 + 1/12 = 1/6$ and so $\text{var}(W) = \sigma^2/16$.

Practice problem

Let X_1, Y_1, X_2, Y_2 be independent random variables, uniformly distributed in the unit interval $[0,1]$. Let

$$W = \frac{X_1 + \cdots + X_{16} - (Y_1 + \cdots + Y_{16})}{16}$$

Find a numerical approximation of $P(|W - E[W]| > .001)$.

- We can rewrite W as $W = \dfrac{(X_1 - Y_1) + (X_2 - Y_2) + \cdots + (X_{16} - Y_{16})}{16}$
- Now note that W is an average of 16 i.i.d random variables. Let $Z_i = X_i - Y_i$.
- $E[Z_i] = E[X_i] - E[Y_i] = 0$ and so $E[W] = 0$.
- $\sigma^2 = \text{var}(Z_i) = \text{var}(X_i - Y_i) = \text{var}(X_i) + \text{var}(Y_i) = 1/12 + 1/12 = 1/6$ and so $\text{var}(W) = \sigma^2/16$.
- Standardize W as $W_s = \dfrac{W - 0}{\sigma/4}$. We know that $P(W_s < c) \approx \Phi(c)$
- So

$$\begin{aligned} P(|W - E[W]| > .0001) &= P(|W| > 0.001) = P(|W|/(\sigma/4)) > 0.001/(\sigma/4)) \\ &= P(|W_s| > 0.001/(\sigma/4)) \approx P(|Z| > .0098) = 0.9922 \end{aligned}$$

Normal Approximation to Binomial

Consider a Binomial random variable $X \sim Binomial(n, p)$. For large n we can calculate the CDF of X by looking up the normal lookup table, which is much easier than evaluating large factorials.

- $X = \sum_i Y_i$, where Y_i are independent Bernoulli(p) random variables. Let $M_n = X/n$.
- $E[M_n] = p$ and $\text{var}(M_n) = p(1-p)/n$
- $P(X \le x) = P\left(M_n \le \frac{x}{n}\right) = P\left(\frac{\sqrt{n}(M_n - p)}{\sqrt{p(1-p)}} \le \underbrace{\frac{\sqrt{n}(x/n - p)}{\sqrt{np(1-p)}}}_{\frac{(x-np)}{\sqrt{p(1-p)}}}\right) \approx \Phi\left(\frac{(x-np)}{\sqrt{p(1-p)}}\right)$
- The DeMoivre Laplace limit theorem states that for $x_2 < x_1$, $P(x_2 \le X \le x_1) \approx \Phi\left(\frac{(x_1 - np)}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{(x_2 - np)}{\sqrt{p(1-p)}}\right)$

Confidence Interval

- If I give you n independent datapoints with $E[X_1] = \cdots = E[X_n] = \mu$
- Also, $var(X_1) = \cdots = var(X_n) = \sigma^2$
- Then $M_n = \sum_i X_i/n$ is a good estimate of μ
- But often we want a confidence interval which shows whether we are confident about our estimator
- We know that M_n is approximately a Gaussian with mean μ and variance σ^2/n
- So we are looking for two numbers that you compute from data u, ℓ such that $P(\mu \in [u, \ell]) = 1 - \alpha$
- Say we could fine $P(|M_n - \mu| \leq t_\alpha) = 1 - \alpha$, then the CI around μ with coverage α will be $[M_n - t_\alpha, M_n + t_\alpha]$

Confidence Interval - known σ

- We know from CLT that $(M_n - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$
- We know how to find z_α such that $P(|Z| \leq z_\alpha) = 1 - \alpha$,
- Then the CI around μ with coverage α will be $[M_n - z_\alpha \sigma/\sqrt{n}, M_n + z_\alpha \sigma/\sqrt{n}]$
- If $\alpha = .05$, then $z_\alpha = 1.96$ (this is also known as the z-score.