EE126: Probability and Random Processes

Lecture 15: Weak Law of Large Numbers

Abhay Parekh

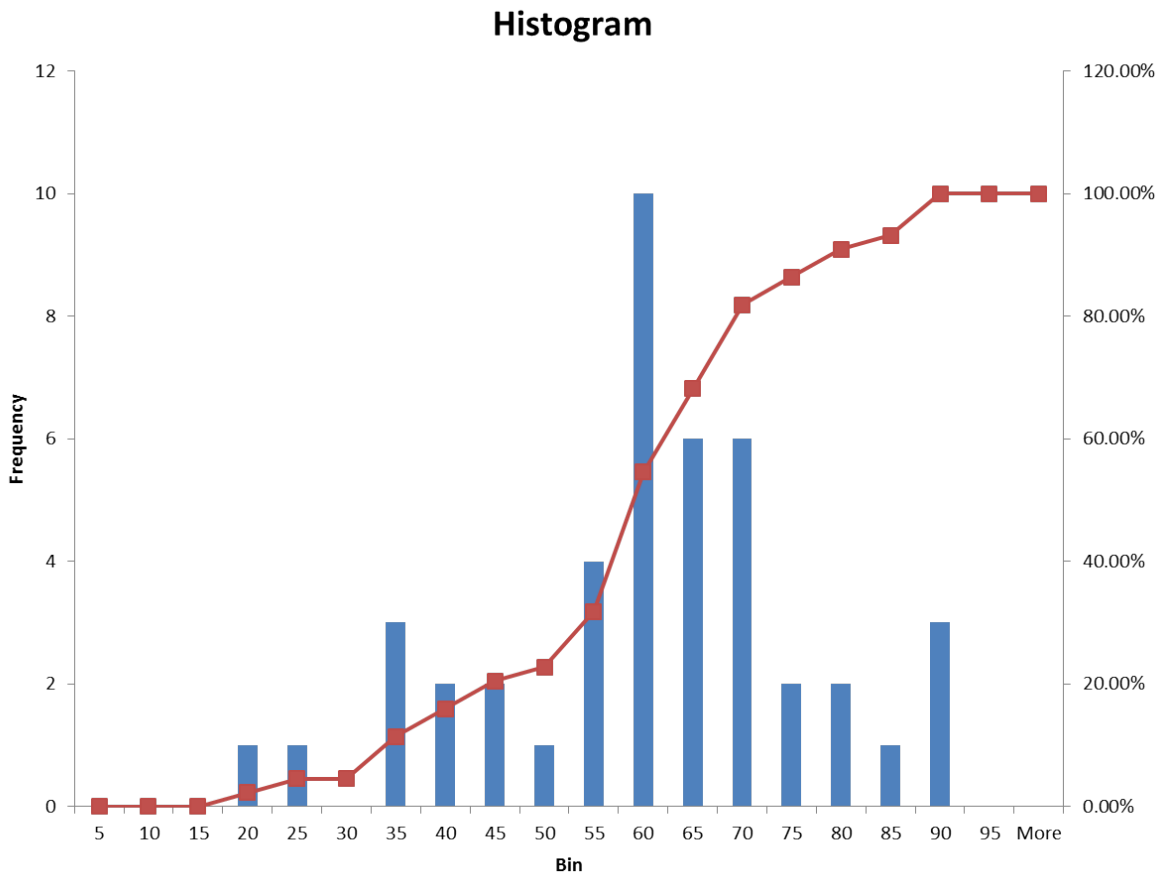
UC Berkeley

March 10, 2011

Logistics

- HW's will now be due end of day Wed instead of Tuesday.
- Midterm
  - It was not an easy exam. You did really well as a group! Most of you should feel very good about your performance.
  - Regrades until Tuesday. See your GSIs or me. I will make final call.
  - Please look at the exam solutions.

Midterm



Mean=58.41, Median = 60 Standard Deviation=16.39.

Your score

≤ 50: Concepts, App
$(50, 70]$: Concepts, App
> 70: Concepts, App.

Review

1. Power of Conditioning: Heavy use of Iterated Expectations Law

$$E[E[X|Y]] = E[X]$$

2. Transforms: Bilateral Laplace Transform

$$M_X(s) = E[e^{sX}] = \begin{cases} \int_x e^{sX} f_X(x)dx, & X \text{ continuous;} \\ \sum_x e^{sX} p_X(x), & X \text{ discrete.} \end{cases}$$

$E[X|Y]$ as an Estimator

Suppose we want to estimate X from an observation $Y = y$. Then

$$E[(X - \hat{X})^2|Y = y] \text{ is minimized at } \hat{X} = E[X|Y = y]$$

1. $E[X|Y]$ is an unbiased estimator.

$$E[\underbrace{X - \hat{X}}_{\text{estimation error}}] = E[X - E[X|Y]] = E[X] - E[X] = 0$$

2. $\hat{X} = E[X|Y]$ is uncorrelated with the estimation error $\hat{X} - X$.

$$cov(\hat{X}, \hat{X} - X) = 0$$

3. $$var(\hat{X} + X - \hat{X}) = var(\hat{X}) + \mathrm{var}(X - \hat{X})$$

So

$$\boxed{var(X) = var(E[X|Y]) + var(X - E[X|Y])}$$

Law of Total Variance

Starting from

$$\boxed{var(X) = var(E[X|Y]) + var(X - E[X|Y])}$$

we derived

Given random variables, X, Y:

$$var(X) = var(E[X|Y]) + E[var(X|Y)]$$

Summing a Random Number of Random Variables

Suppose $Y = X_1 + \ldots X_N$, the X_i are iid, but N is a random variable independent of the X_i's. What are $E[Y]$ and $var(Y)$?

$$\boxed{E[Y] = E[N]E[X_i]}$$

$$\boxed{var(Y) = E[X_i]^2 var(N) + E[N]var(X_i)}$$

Transforms

Definition

Given a random variable X, the Transform of X, $M_X(s)$ is defined as

$$M_X(s) = E[e^{sX}]$$

for all scalars s

Finding Moments

For $M_X(s) = E[e^{sx}]$:

$$\frac{d^n M_X(s)}{ds^n}\Big|_{s=0} = E[X^n]$$

Properties:

1. $M_X(0) = 1$
2. If $X > 0$, $M_X(-\infty) = 0$ and if $X < 0$ then $M_X(\infty) = 0$.
3. If $Y = aX + b$ than
   $M_Y(s) = E[e^{s(aX+b)}] = e^{sb}E[e^{aX}] = e^{sb}M_X(as)$

Combinations of Transform Functions

1. Mixture of distributions: Suppose $\sum_{i=1}^{n} p_i = 1$, and $f_X(x) = \sum_{i=1}^{n} p_i f_{X_i}(x)$. Then

$$\boxed{M_X(s) = \sum_{i=1}^{n} p_i M_{X_i}(s)}$$

2. Sum of Independent Random Variables: $Z = X + Y$; X, Y independent. Then

$$M_Z(s) = E[e^{(X+Y)s}] = E[e^{Xs}e^{Ys}] = E[e^{Xs}]E[e^{Ys}] = M_X(s)M_Y(s)$$

So convolving the densities corresponds to multiplying transforms.

Example

If X_i is bernoulli with with parameter p then $M_{X_i} = 1 - p + pe^s$ for $i = 1, 2, ..., n$.
$Y = \sum_i X_i$ is a Binomial Random Variable.

$$M_Y(s) = \Pi_{i=1}^{n}(1 - p + pe^s) = (1 - p + pe^s)^n.$$

$$E[X] = n(1 - p + pe^s)^{n-1}pe^s\Big|_{s=0} = n(1)^{n-1}p = np$$

$$\begin{aligned} E[X^2] &= np[(n-1)(1 - p + pe^s)^{n-2}pe^{2s} + (1 - p + pe^s)^{n-1}e^s]\Big|_{s=0} \\ &= np(1 - p + np). \end{aligned}$$

Transform of Sum of Random Number of RVs

Let $Y = X_1 + ... + X_N$ where $X_i, i = 1, 2, ..., n$ are iid and N is a random variable.
To find $M_Y(s)$:

1. Find $M_N(s)$
2. Replace s with $\ln M_X(s)$, i.e. e^s **with** $M_X(s)$**.**

Example:
Each of 3 gas station is open on any given day with prob $\frac{1}{2}$
The amount of gas available is uniformly distributed on $[0, 1000]$.
Let Y be the total amount of gas available on any given day. Find $M_Y(s)$.
N: number of gas stations open:
$M_N(n) = (1 - 0.5 + 0.5e^s)^3 = \frac{1}{8}(1 + e^s)^3$.
Now

$$M_X(s) = \frac{e^{1000s} - 1}{1000s}$$

(Look this up)
So

$$M_Y(s) = \frac{1}{8}(1 + \frac{e^{1000s} - 1}{1000s})^3$$

Limiting Behavior of Random Variables

The basic framework for probability requires doing an "experiment"
What happens if we do this experiment many many times?
How do the probabilities behave relative to the observed frequencies?
Example: Let $X_i = 1$ if we observe some event in the sample space, $X_i = 0$ if we do not observe it. Then $M_n = \frac{\sum_{i=1}^{n} X_i}{n}$ is the fraction of time we observe X_i in n trials. The trials are independent and so are the X_i. We know that:

1. $E[M_n] = \frac{nE[X_i]}{n} = E[X_i]$
2. Assuming $var(X_i)$ exists, $var(M_n) = \frac{n\ var(X_i)}{n^2} = \frac{var(X_i)}{n}$

As $n \to \infty$: $E[M_n] = E[X_i]$ and $var(M_n) = 0$.
The bulk of the probability of M_n becomes concentrated at $E[X_i]$!
This is true of any random variable X!

Inequalities

The fact that the variance $\to \infty$ at rate $\frac{1}{n}$ is great but what does that tell us about $P(|M_n - E[X_i|)$?

1. How quickly does it go to zero?
2. What happens if the variance is infinite?
3. Is it always OK to think of probabilities as relative frequencies?

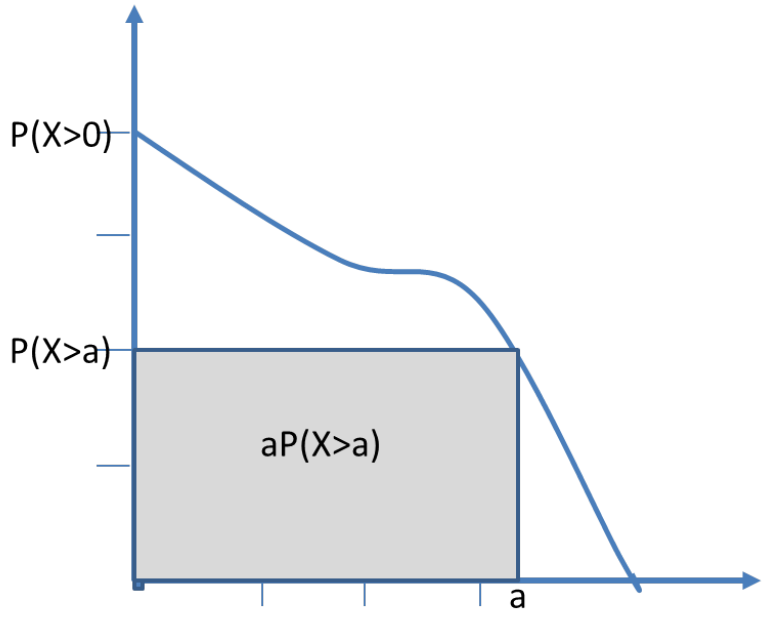
To figure this out, we will prove some inequalities which are cool and important in their own right!

Markov Inequality

If X can only take non-negative values then

$$P(X \geq a) \leq \frac{E[X]}{a}$$

for all $a > 0$.



Markov Inequality

If X can only take non-negative values then

$$P(X \geq a) \leq \frac{E[X]}{a}$$

for all $a > 0$.

This inequality makes no assumptions on the existence of variance and so it can't be very strong for typical distributions. In fact, it is quite weak.

Markov Inequality

Example: X is the height of a random adult in Berkeley. If $E[X] = 68$ inches, the Markov Inequality says that

$$P(X > 144) \leq \frac{68}{144} = 0.47$$

On the other hand since it is general, we can try to see what happens to it as we add more assumptions on the distribution. Think of this inequality as being the building block for others...

Chebyshev Inequality

Suppose X has finite variance σ^2. Then let $Y = (X - E[X])^2$, and since $Y \geq 0$ apply the Markov Inequality to it:

$$\begin{aligned} P(Y \geq a) &\leq \frac{E[Y]}{a} \\ P((X - E[X])^2 \geq a) &\leq \frac{E[(X - E[X])^2}{a} = \frac{\sigma^2}{a} \\ P(|X - E[X]| \geq \sqrt{a}) &\leq \frac{\sigma^2}{a} \\ P(|X - E[X]| \geq c) &\leq \frac{\sigma^2}{c^2} \end{aligned}$$

This is the famous Chebyshev Inequality

Chebyshev Inequality

If X is a random variable with finite mean and variance σ^2, then

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

for all $c > 0$.
Also, letting $c = k\sigma$:

$$P(|X - E[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

Now we connect variance to probability...If $var(X) = 0$ then

$$P(|X - E[X]| \geq c) = 0$$

for all c.

Chebyshev Inequality

Example: We want to estimate $E[X]$ by measuring random adults and computing the average height (M_n). How many should we measure to ensure that the estimate is within 1 inch of $E[X]$ with probability 0.99? Assume that $\sigma_X^2 = 36$.
$var(M_n) = \frac{36}{n}$. We want

$$P(|M_n - E[X]| \leq 1) = 1 - P(|M_n - E[X]| > 1) \geq 0.99$$

$$P(|M_n - E[X]| > 1) \leq 0.01$$

$E[M_n] = E[X]$, so Chebyshev tells us that

$$P(|M_n - E[X]| \geq 1) \leq \frac{36}{n}$$

$$\frac{36}{n} \leq .01 \Rightarrow n \geq 3600.$$

In general to have 0.99 confidence: $n \geq \frac{100\sigma^2}{c^2}$.

Example: Bounded Range Random Variable

Let X take values only in $[a, b]$. Suppose it has finite variance, σ^2. *Find a bound on $P(|X - E[X]| \geq c)$ for all c that is independent of σ^2).*

From last lecture we know that $E[(X - \hat{X})^2]$ is minimized at $\hat{X} = E[X]$. So, for any constant γ:

$$\sigma^2 = E[(X - \hat{X})^2] \leq E[(X - \gamma)^2].$$

Ideally, we would like to find a γ in terms of a and b for which the RHS is the smallest.

We will show that

$$P(|X - E[X]| \geq c) \leq \frac{(b-a)^2}{4c^2}$$

and that this is the best possible upper bound.

Example: Bounded Range Random Variable

$$\begin{aligned}
\sigma^2 &= E[(X-\hat{X})^2] \leq E[(X-\gamma)^2]. \\
&\quad \text{Pick } \gamma = \tfrac{a+b}{2} \text{ to get} \\
&\leq E[(X-\frac{a+b}{2})^2] = E[X^2 - X\frac{a+b}{2} + \frac{(a+b)^2}{4}] \\
&= E[X^2 - aX - bX + \frac{(a+b)^2}{4} - X\frac{a+b}{4} + ab + \frac{(b-a)^2}{4}] \\
&= E[(X-a)(X-b)] + \frac{(b-a)^2}{4}] \leq \frac{(b-a)^2}{4}
\end{aligned}$$

Inequality is tight, i.e. let $P(X=a) = P(X=b) = 0.5$ then $\sigma^2 = \frac{(b-a)^2}{4}$.
Now substitute in Chebyshev:

$$\boxed{P(|X-E[X]| \geq c) \leq \frac{\sigma^2}{c^2} \leq \frac{(b-a)^2}{4c^2}}$$

Chebyshev: Another derivation

$$E[X^2] = \int_x x^2 f_X(x)dx \geq \int_{x:|x|\geq\epsilon} x^2 f_X(x)dx$$

Now,

$$\int_{x:|x|\geq\epsilon} x^2 f_X(x)dx \geq \epsilon^2 \int_{x:|x|\geq\epsilon} f_X(x)dx$$

So

$$E[X^2] \geq \epsilon^2 P(|X| \geq \epsilon)$$

I.e.,

$$P(|X| \geq \epsilon) \leq \frac{E[X^2]}{\epsilon^2}$$

(To get back the original form just substitute $X = Y - E[Y]$.)
To get equality for a given ϵ, define X such that $P(X = \epsilon) = 1$.

Example: Random Walk

Starting at the origin, I flip a coin $10,000$ times: on each flip, if it is heads I step forward; if it is tails I step backward. Prob of heads is 0.5. How far should I expect to be from the origin?

For now let's say the coin is tossed n times. Let $X_i = 1$ if toss i is heads and -1 otherwise.

$E[X_i] = 0$ and $E[X_i^2] = 1$ so $var(X_i) = 1$

Let $X = \sum_i X_i$. $E[X] = 0$ and $var(X) = n$. But this isn't helpful because $+$ive outcomes are reduced by -ive outcomes. We want $|X|$ since that's the magnitude of the distance.

Chebyshev says:

$$P(|X| \geq k\sqrt{n}) \leq \frac{1}{k^2}$$

$$P(|X| \geq 100k) \leq \frac{1}{k^2}$$

So for example, the prob I am more than 400 steps away is less than $\frac{1}{16}$.

Chernoff Bound

Also builds on Markov: $P(X \geq a) \leq \frac{E[X]}{a}$ for $a > 0$.
Pick $X = e^{Ys}$. Then

$$P(e^{Ys} \geq a) \leq \frac{M_Y(s)}{a}$$

Now let $a = e^{sb}$. Then for $s > 0$:

$$P(Y \geq b) \leq e^{-sb} M_Y(s)$$

and for $s < 0$:

$$P(Y \leq b) \leq e^{-sb} M_Y(s)$$

Note that the LHS does not depend on s (other than its sign) so we can optimize the RHS to get the best bound.
Since $M_Y(s)$ has all the information contained in $f_Y(y)$ if the distribution has tails that drop off sharply, that will be reflected in our bound.

Cheroff Bound

Example: for a standard normal distribution: $M_Y(S) = e^{-\frac{y^2}{2}}$. It turns out that applying the Chernoff bound gives:

$$P(Y \geq b) \leq e^{\frac{-b^2}{2}}$$

So, if height, X is a normal rv with mean 68 and variance 36, $\frac{X-68}{6}$ is a standard normal.

$$P(X \geq 144) = P(\frac{X-68}{6} \geq \frac{76}{36} = 2.11 \leq e^{-\frac{2.11^2}{2}} \approx 0.1077.$$

Weak Law of Large Numbers

Recall: We perform an experiment n times independently and

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

The fact that $var(M_n) \to \infty$ at rate $\frac{1}{n}$ is great but what does that tell us about $P(|M_n - E[X_i|)$? How quickly does it go to zero?
Just use Chebyshev: $P(|X - E[X]| \geq c) \leq \frac{\sigma^2}{c^2}$

$$P(|M_n - E[X_i]) \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

for any $\epsilon > 0$.
This tells us what we suspected. This is a form of the Weak Law of Large Numbers.

Weak Law of Large Numbers

It turns out that the WL also holds when variances are infinite, but we can't prove that here.

Weak Law of Large Numbers

If $X_1, ..., X_n$ are iid random variables with mean μ then for every $\epsilon > 0$:

$$P(|\frac{X_1 + \ldots X_n}{n} - \mu| \geq \epsilon) \to 0, \quad \text{as } n \to \infty.$$

This tells us that empirical frequencies are good estimates of p.