Expectation is linear

- So far we saw that $E(X+Y) = E(X) + E(Y)$.
- Let $\alpha \in \mathbb{R}$. Then,

$$\begin{aligned} E(\alpha X) &= \sum_{\omega} (\alpha X)(\omega) \Pr(\omega) \\ &= \sum_{\omega} \alpha X(\omega) \Pr(\omega) \\ &= \alpha \sum_{\omega} X(\omega) \Pr(\omega) \\ &= \alpha E(X). \end{aligned}$$

- **Corollary.** For $\alpha, \beta \in \mathbb{R}$,

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y).$$

Expectation of $\varphi(X)$

- X is a random variable and $\varphi : \mathbb{R} \mapsto \mathbb{R}$.
- We want the expectation of $Y = \varphi(X)$.
- We can compute
  $f_Y(y) = \Pr(\varphi(X) = y) = \Pr(\{\omega : X(\omega) \in \varphi^{-1}(y)\})$,
  and use $E(Y) = \sum_{y \in \mathcal{R}_Y} y f_Y(y)$, where $\mathcal{R}_Y$ is the range of Y.
- Alternatively we have,
  Claim. $E(\varphi(X)) = \sum_{x \in \mathcal{R}_X} \varphi(x) f_X(x)$.
  Proof.
  $$\begin{aligned} E(\varphi(X)) &= \sum_{\omega} \varphi(X(\omega)) \Pr(\omega) \\ &= \sum_{x \in \mathcal{R}_X} \sum_{\omega : X(\omega) = x} \varphi(X(\omega)) \Pr(\omega) \\ &= \sum_{x} \sum_{\omega : X(\omega) = x} \varphi(x) \Pr(\omega) \\ &= \sum_{x} \varphi(x) f_X(x). \end{aligned}$$
- **Example.** For a random variable X,
  $$E(X^2) = \sum_{x} x^2 f_X(x).$$

Variance of X

- Consider the following three distributions:

$$f_X(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} 1/2 & y = -1, 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Z(z) = \begin{cases} 1/2 & z = -100, 100 \\ 0 & \text{otherwise} \end{cases}$$

- What are the expectations of these distributions?
- Does the expectation tell the "whole story"?
- Clearly Z is much more spread about its mean than X and Y.
- An intuitively appealing measurement of the spread of X about its mean $\mu = E(X)$ is given by $E(|X - \mu|)$.
- **Def.** For convenience the *variance* of X is defined as

$$V(X) = E(X - \mu)^2.$$

- **Def.** The *standard deviation* is $\sigma(X) = \sqrt{V(X)}$.

Examples

- Let X be Bernoulli(p). We saw that $\mu = p$.

$$\begin{aligned} V(X) &= (0-p)^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p(1-p)[p + (1-p)] \\ &= p(1-p). \end{aligned}$$

- **Claim.** $V(X) = E(X^2) - \mu^2$.
  Proof.

$$\begin{aligned} E(X-\mu)^2 &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + E(\mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

- X is the outcome of a roll of a fair die.
  - · We saw that $E(X) = 7/2$.
  - · $E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \cdots + 6^2 \cdot \frac{1}{6} = \frac{91}{6}$.
  - · So, $V(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$.

$V(X+Y)$

- Let X and Y be random variables with $\mu = E(X)$ and $\nu = E(Y)$.
- **Def.** The *covariance* of X and Y is

$$\text{Cov}(X,Y) = E(XY) - E(X) \cdot E(Y).$$

- **Claim.** $V(X+Y) = V(X) + V(Y) + 2 \cdot \text{Cov}(X,Y)$.
- **Proof.** $E(X+Y) = \mu + \nu$, so

$$\begin{aligned} V(X+Y) &= E[(X+Y)^2] - (\mu+\nu)^2 \\ &= E(X^2 + 2XY + Y^2) - (\mu^2 + 2\mu\nu + \nu^2) \\ &= [E(X^2) - \mu^2] + [E(Y^2) - \nu^2] \\ &\qquad + 2 \cdot [E(XY) - \mu\nu] \end{aligned}$$

Suppose X and Y are independent

- **Claim.** If X and Y are independent $\mathrm{Cov}(X, Y) = 0$.
- **Proof.**

$$
\begin{aligned}
E(XY) &= \sum_{\omega} (XY)(\omega) \Pr(\omega) \\
&= \sum_{x \in \mathcal{R}_X} \sum_{y \in \mathcal{R}_Y} \sum_{\omega : X(\omega) = x \,,\, Y(\omega) = y} X(\omega) \cdot Y(\omega) \cdot \Pr(\omega) \\
&= \sum_{x} \sum_{y} \sum_{\omega : X(\omega) = x \,,\, Y(\omega) = y} x \cdot y \cdot \Pr(\omega) \\
&= \sum_{x} \sum_{y} x \cdot y \cdot \Pr(X = x, Y = y) \\
&= \sum_{x} \sum_{y} x \cdot y \cdot \Pr(X = x) \cdot \Pr(Y = y) \\
&= \sum_{x} x \cdot \Pr(X = x) \sum_{y} y \cdot \Pr(Y = y) \\
&= E(X) \cdot E(Y).
\end{aligned}
$$

- **Corollary.** If X and Y are independent

$$V(X + Y) = V(X) + V(Y).$$

The variance of $B_{n,p}$

- **Corollary.** If $X_1, \dots X_n$ are independent then $V(X_1+X_2+\dots+X_n) = V(X_1)+V(X_2)+\dots V(X_n)$.
  Proof. By induction but note that we need to show that $X_1 + \dots + X_{k-1}$ is independent of X_k.
- Let X be a $B_{n,p}$ random variable.
- Then $X = \sum_1^n X_k$ where X_k are independent Bernoulli p random variables. So,
$$V(X) = V\Big(\sum_1^n X_k\Big) = \sum_1^n V(X_k) = np(1-p).$$
- For a fixed p the variance increases with n.
- Does this make sense?
- For a fixed n the variance is minimized for $p = 0, 1$ and maximized for $p = 1/2$.
- Does it make sense?
- Expectation and variance are just two "measurements" of the distribution. They cannot possibly convey the same amount of information that is in the distribution function.
- Nevertheless we can learn a lot from them.

Markov's Inequality

- **Theorem.** Suppose X is a nonnegative random variable and $\alpha > 0$. Then

$$\Pr(X \geq \alpha) \leq \frac{E(X)}{\alpha}.$$

- **Proof.**

$$\begin{aligned} E(X) &= \sum_x x \cdot f_X(x) \\ &\geq \sum_{x \geq \alpha} x \cdot f_X(x) \\ &\geq \sum_{x \geq \alpha} \alpha \cdot f_X(x) \\ &= \alpha \sum_{x \geq \alpha} f_X(x) \\ &= \alpha \cdot \Pr(X \geq \alpha). \end{aligned}$$

- **Example.** If X is $B_{100,1/2}$,

$$\Pr(X \geq 100) \leq \frac{50}{100}.$$

This is not very accurate: the correct answer is ... $2^{-100} \sim 10^{-30}$.

- What would happen if you try to estimate this way $\Pr(X \geq 49)$?

Chebyshev's Inequality

- **Theorem.** X is a random variable and $\beta > 0$.

$$\Pr(|X - \mu| \geq \beta) \leq \frac{V(X)}{\beta^2}.$$

- **Proof.** Let $Y = (X - \mu)^2$. Then,

$$|X - \mu| \geq \beta \iff Y \geq \beta^2,$$

So

$$\{\omega : |X(\omega) - \mu| \geq \beta\} = \{\omega : Y(\omega) \geq \beta^2\}.$$

In particular, the probabilities of these events are the same:

$$\Pr(|X - \mu| \geq \beta) = \Pr(Y \geq \beta^2).$$

Since $Y \geq 0$ by Markov's inequality

$$\Pr(Y \geq \beta^2) \leq \frac{E(Y)}{\beta^2}.$$

Finally, note that $E(Y) = E[(X - \mu)^2] = V(X)$.

Example

- Chebyshev's inequality gives a lower bound on how well is X concentrated about its mean.
- Suppose X is $B_{100,1/2}$ and we want a lower bound on $\Pr(40 < X < 60)$.
- Note that

$$\begin{aligned} 40 < X < 60 &\iff -10 < X - 50 < 10 \\ &\iff |X - 50| < 10 \end{aligned}$$

so,

$$\begin{aligned} \Pr(40 < X < 60) &= \Pr(|X - 50| < 10) \\ &= 1 - \Pr(|X - 50| \geq 10). \end{aligned}$$

Now,

$$\begin{aligned} \Pr(|X - 50| \geq 10) &\leq \frac{V(X)}{10^2} \\ &= \frac{100 \cdot (1/2)^2}{100} \\ &= \frac{1}{4}. \end{aligned}$$

So,

$$\Pr(40 < X < 60) \geq 1 - \frac{V(X)}{10^2} = \frac{3}{4}.$$

- This is not too bad: the correct answer is ~ 0.9611.

The law of large numbers (LLN)

- You suspect the coin you are betting on is biased.
- You would like to get an idea on the probability that it lands heads. How would you do that?
- Flip n times and check the relative number of Hs.
- In other words, if X_k is the indicator of H on the kth flip, you estimate p as
$$p \approx \frac{\sum_{k=1}^{n} X_k}{n}.$$
- The underlying assumption is that as n grows bigger the approximation is more likely to be accurate.
- Is there a mathematical justification for this intuition?

LLN cont.

- Consider the following betting scheme:
  - At every round the croupier rolls a die.
  - You pay \$1 to join the game in which you bet on the result of the next 5 rolls.
  - If you guess them all correctly you get $6^5 = 7776$ dollars, 0 otherwise.
  - How can you estimate if this is a fair game?
  - Study the average winnings of the last n gamblers.
- Formally, let X_k be the winnings of the kth gambler.
- We hope to estimate $E(X_k)$ by

$$E(X_k) \approx \frac{\sum_{k=1}^{n} X_k}{n}.$$

- Is there a mathematical justification for this intuition?
- Is the previous problem essentially different than this one?

Example of the (weak) LLN

Consider again the binomial $p = 1/2$ case. With

$$S_n = \sum_{k=1}^{n} X_k,$$

we expect, for example, that

$$\Pr(0.4 < \frac{S_n}{n} < 0.6) = \Pr(0.4n < S_n < 0.6n)$$

will be big (close to 1) as n increases.
As before,

$$\begin{aligned}\Pr(0.4n < S_n < 0.6n) &= \Pr(-0.1n < S_n - 0.5n < 0.1n) \\ &= \Pr(|S_n - 0.5n| < 0.1n) \\ &= 1 - \Pr(|S_n - 0.5n| \geq 0.1n).\end{aligned}$$

As before we can bound

$$\begin{aligned}\Pr(|S_n - 0.5n| \geq 0.1n) &\leq \frac{V(S_n)}{(0.1n)^2} \\ &= \frac{n \cdot (1/2)^2}{0.01n^2} \\ &= \frac{1}{0.04n}.\end{aligned}$$

$$\Rightarrow \Pr(0.4 < \frac{S_n}{n} < 0.6) \geq 1 - \frac{1}{0.04n} \xrightarrow[n\to\infty]{} 1.$$

Are any of 0.4, 0.6 or $p = 1/2$ special?

The (weak) law of large numbers

- The previous example can be generalized to the following statement about a sequence of Bernoulli(p) trials: for any $\varepsilon > 0$,

$$\Pr\left(\left|\frac{\sum_{k=1}^{n} X_k}{n} - p\right| \geq \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

- A further generalization allows us to replace p by $E(X_k)$.

- Suppose $X_1, X_2, \ldots$ are a sequence of iid (independent and identically distributed) random variables. Then, with $\mu = E(X_k)$

$$\Pr\left(\left|\frac{\sum_{k=1}^{n} X_k}{n} - \mu\right| \geq \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

- The proof is essentially identical to the previous one using Chebyshev's inequality.

The binomial dispersion

- S_n is a binomial $B_{(n,p)}$ random variable.
- How tightly is it concentrated about its mean?
- In particular, how large an interval about the mean should we consider in order to guarantee that S_n is in that interval with probability of at least 0.99?
- Can you readily name such an interval?
- Can we be more frugal?
- We know that if we take an interval of length, say, $2 \cdot n/10$ then

$$\Pr(np - n/10 < S_n < np + n/10) \xrightarrow[n\to\infty]{} 1.$$

- Why is that true?

$$\begin{aligned} & np - n/10 < S_n < np + n/10 \\ \iff & -1/10 < \frac{S_n - np}{n} < 1/10 \\ \iff & \left|\frac{S_n}{n} - p\right| < 1/10, \end{aligned}$$

and by the LLN

$$\Pr\left(\left|\frac{S_n}{n} - p\right| < 1/10\right) \xrightarrow[n\to\infty]{} 1.$$

- Therefore, for all "sufficiently large" n,

$$\Pr(np - n/10 < S_n < np + n/10) \leq 0.99.$$

- Two problems:
    · We didn't really say what n is?
    · We are still being "wasteful" (as you will see).
- Clearly, the question is that of the dispersion of S_n about its mean.
- Recall that the variance is supposed to (crudely) measure just that.
- Chebyshev's inequality helps visualizing that: $\forall \beta > 0$

$$\begin{aligned}\Pr(|X - E(X)| < \beta) &= 1 - \Pr(|X - E(X)| \geq \beta) \\ &\geq 1 - \frac{V(X)}{\beta^2}.\end{aligned}$$

- What should β be in order to make sure that

$$\Pr\Big(|X - E(X)| < \beta\Big) \geq 0.99\ ?$$

- Need: $\frac{V(X)}{\beta^2} \leq 0.01$, or

$$\beta \geq \sqrt{100V(X)} = 10\sigma(X).$$

- Applying this general rule to the binomial S_n we have

$$\mathrm{P}\left(|S_n - np| < 10\sqrt{np(1-p)}\right) \geq 0.99.$$

- More generally,

$$\Pr\Big(|S_n - np| < \alpha\sigma(S_n)\Big) \geq 1 - \frac{1}{\alpha^2}.$$

- Since $\sigma(S_n) = \sqrt{np(1-p)}$, it means most of "the action" takes place in an interval of size $c\sqrt{n}$ about np (before we had an interval of size $b \cdot n$).

Confidence interval

- What happens if we don't know p?
- We can still repeat the argument above to get:

$$\begin{aligned}\Pr(|S_n - np| < 5\sqrt{n}) &\geq 1 - \frac{V(S_n)}{(5\sqrt{n})^2} \\ &= 1 - \frac{np(1-p)}{25n} \\ &\geq 1 - \frac{1/4}{25} = 0.99,\end{aligned}$$

since $p(1-p) \leq 1/4$.

- It follows that for any p and n:

$$\Pr\left(\left|\frac{S_n}{n} - p\right| < \frac{5}{\sqrt{n}}\right) \geq 0.99.$$

- So with probability of at least 0.99, S_n/n is within a distance of $5/\sqrt{n}$ of its *unknown* mean, p.
- This can help us design an experiment to estimate p.
- For example, suppose that a coin is flipped 2500 times and that $S = S_{2500}$ is the number of heads.
- Then with probability of at least 0.99, $S/2500$ is within $5/50 = 0.1$ of p.

- Equivalently, with probability of at least 0.99 the interval $(S/2500 - 0.1, S/2500 + 0.1)$ contains p.
- Such an interval is called a 99% confidence interval for p.
- For example, suppose we see only 750 heads in 2500 flips.
- Since $750/2500 = 0.3$ our 99% confidence interval is $(0.3 - 0.1, 0.3 + 0.1) = (0.2, 0.4)$.
- We should therefore be quite suspicious of this coin.
- **Remark.** We have been quite careless: all we used to generate our confidence interval was Chebyshev's inequality. Chebyshev's inequality doesn't "know" that S_n happens to be a binomial random variable: it only uses the mean and the variance of S_n. A more careful analysis would gain us a significantly tighter 99% confidence interval.