MATH 2030 3.00MW – Elementary Probability Course Notes Part V: Independence of Random Variables, Law of Large Numbers, Central Limit Theorem, Poisson distribution Geometric & Exponential distributions

Tom Salisbury – salt@yorku.ca

York University, Dept. of Mathematics and Statistics

Original version April 2010. Thanks are due to
E. Brettler, V. Michkine, and R. Shieh for many corrections

January 19, 2018

Independence

- Random variables X, Y are *independent* if
  $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$
  for all choices of $A, B \subset \mathbb{R}$.
- So knowing the value of X gives no information about Y.
- We'll generally stick to 2 r.v., but everything generalizes to more. Eg. $X_1, \ldots, X_n$ are *independent* if
  $P(X_i \in A_i, i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i \in A_i)$.
- If X, Y are discrete then independence $\Leftrightarrow$
  $P(X = x, Y = y) = P(X = x)P(Y = y)$ for every x, y.

  [Pf: For $\Rightarrow$ take $A = \{x\}$, $B = \{y\}$. For $\Leftarrow$,
  $P(X \in A)P(Y \in B) = (\sum_{x \in A} P(X = x))(\sum_{y \in B} P(Y = y))$
  $= \sum_{x \in A, y \in B} P(X = x)P(Y = y)$
  $= \sum_{x \in A, y \in B} P(X = x, Y = y) = P(X \in A, Y \in B)$]
- There's a similar result in the continuous case, but it uses *joint densities*, something we aren't going to get into.

Independence

Some more basic properties of independence:

- X, Y independent $\Rightarrow g(X), h(Y)$ independent, $\forall g, h$.

  [Proof: Fix A, B and let $C = \{x \mid g(x) \in A\}$,
  $D = \{y \mid h(y) \in B\}$. Then
  $P(g(X) \in C, h(Y) \in D) = P(X \in C, Y \in D)$
  $= P(X \in C)P(Y \in D) = P(g(X) \in A)P(h(Y) \in B)$.]

- X, Y independent $\Rightarrow E[XY] = E[X]E[Y]$
  (assuming X, Y, XY are integrable).

  [Proof: We'll only do the discrete case.
  $E[X]E[Y] = \left(\sum_x xP(X = x)\right)\left(\sum_y yP(Y = y)\right)$
  $= \sum_{x,y} xyP(X = x)P(Y = y) = \sum_{x,y} xyP(X = x, Y = y)$
  $= E[XY]$. To see the latter, we could use that
  $XY = \sum_{x,y} xy1_{\{X=x, Y=y\}}$. The latter holds because for any
  ω, the only term of the sum that isn't $= 0$ is the one for the x
  and y such that $X(\omega)Y(\omega) = xy$.]

Independence

- X, Y independent $\Rightarrow \text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$

  [Proof: $\text{Var}[X+Y] = E[(X+Y-E[X+Y])^2]$
  $= E\left[\Big((X-E[X]) + (Y-E[Y])\Big)^2\right]$
  $= E\left[(X-E[X])^2 + 2(X-E[X])(Y-E[Y]) + (Y-E[Y])^2\right]$
  $=$
  $E[(X-E[X])^2] + 2E[(X-E[X])(Y-E[Y])] + E[(Y-E[Y])^2]$
  claim 2nd term $=0$, so this $= \text{Var}[X] + \text{Var}[Y]$.
  To see this, note that the factors are independent (a previous property), so by another previous property, it
  $= 2E[X-E[X]]E[Y-E[Y]] = 2(E[X]-E[X])(E[Y]-E[Y])$
  $= 2 \times 0 \times 0 = 0$.]

Calculating via independence

(modified from what was done in class)
Eg: Suppose X and Y are independent, X has mean 3 and variance 5, Y has mean -2 and variance 4.

- Find $E[(X+Y)(2X+3Y+6)]$.
  Solution: $= 2E[X^2] + 5E[XY] + 3E[Y^2] + 6E[X] + 6E[Y]$.
  But $E[X^2] = \text{Var}[X] + E[X]^2 = 5 + 3^2 = 14$,
  $E[Y]^2 = \text{Var}[Y] + E[Y]^2 = 4 + (-2)^2 = 8$, and
  $E[XY] = E[X]E[Y] = 3(-2) = -6$ by independence.
  Putting it all together we get an answer of 28.
- Find $\text{Var}[XY]$
  Solution: $= E[(XY)^2] - E[XY]^2 = E[X^2Y^2] - (E[X]E[Y])^2$
  $= E[X^2]E[Y^2] - E[X]^2E[Y]^2$ by independence. Using the numbers calculated above, $= 14 \times 8 - 3^2(-2)^2 = 76$.

Law of Large Numbers (LOLN)

- Let $X_1, X_2, \ldots, X_n$ be independent, with identical distributions (we say they're *IID*), and mean μ. Let $S_n = X_1 + \cdots + X_n$.
- The *Law of Large Numbers* says, heuristically, that $S_n/n \approx \mu$.
- Simple reason: If $\sigma^2 = \text{Var}[X_1]$ then
  $\text{Var}[S_n/n] = n\sigma^2/n^2 = \sigma^2/n \to 0$,
  so S_n/n is close to its mean, which is μ.
- This explains the long-run-average interpretation of expectations, which says that if X_k denotes our winnings from the k'th round of some game, then in the long run, the amount we win per round $(= S_n/n)$ is μ.
- It also explains our frequentist interpretation of probabilities: If A_k is an event arising from the k'th repeated trial of some experiment, then the relative frequency with which the A's occur is $\frac{1}{n} \sum 1_{A_k}$. And this is $\approx E[1_{A_1}] = P(A_1)$.

LOLN

- All those conclusions come out of the heuristic $S_n/n \approx \mu$. But mathematically, we need a more precise formulation. This'll be the idea that the probability of finding S_n/n "significantly" deviating from μ gets small as n gets large.
- **Thm (LOLN)** $X_1, X_2, \ldots$ IID integrable r.v., mean μ. Then $\forall \epsilon > 0$, $P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0$ as $n \to \infty$.
- We prove this assuming $\sigma^2 = \mathrm{Var}[X] < \infty$. The proof uses **Chebyshev's inequality:** Assume $E[Y] = \mu$, $a > 0$. Then $P\left(\left|Y - \mu\right| \geq a\right) \leq \frac{\mathrm{Var}[Y]}{a^2}$.
- Variant: Let $\sigma^2 = \mathrm{Var}[Y]$. Then $P\left(\left|Y - \mu\right| \geq b\sigma\right) \leq \frac{1}{b^2}$.
- Proof of LOLN: $P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}[S_n/n]}{\epsilon^2} = \frac{\mathrm{Var}[S_n]}{n^2\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$

Chebyshev's ineq.

- The point of Chebyshev is that it gives a universal bound on the probabilities of deviations from the mean – you don't need to know much about the distribution of Y in order to apply it.
- Something this general won't necessarily give a very tight bound in special cases. For example, deviation probabilities for the normal distribution decay much faster than what Chebyshev gives.
- **Proof of Chebyshev:**
  If $|Y - \mu| \geq a$ then $\frac{(Y-\mu)^2}{a^2} \geq 1$. So $1_{\{|Y-\mu| \geq a\}} \leq \frac{(Y-\mu)^2}{a^2}$. Chebyshev now follows by taking expectations.,

Markov's inequality

If $X \geq 0$ and $a > 0 \Rightarrow P(X \geq a) \leq \dfrac{E[X]}{a}$.

- Like Chebyshev, this gives a bound on probabilities, but now for prob's of large values, not values far from the mean
- The bound only depends on the mean of X (but also applies only to positive random variables)
- **Proof:** I'll do it only in the case where there's a density f.
  Then $P(X \geq a) = \int_a^\infty f(x)\,dx$
  $\leq \int_a^\infty \frac{x}{a} f(x)\,dx$ (as $\frac{x}{a} \geq 1$ when $x \geq a$).
  $\leq \int_0^\infty \frac{x}{a} f(x)\,dx = E[X]/a$ (as $X \geq 0$).
- **Eg:** If $X \geq 0$ has mean 2, then Markov $\Rightarrow P(X \geq 4) \leq \frac{1}{2}$.
- If we add the information that $\mathrm{Var}[X] = 1$, then we can improve this: Chebyshev $\Rightarrow P(X \geq 4) \leq P(|X - 2| \geq 2) \leq \frac{1}{4}$.
- And if $X \sim \mathrm{Bin}(4, \frac{1}{2})$ then actually $P(X \geq 4) = \frac{1}{16}$.

Central Limit Theorem

- LOLN says that if $S_n = X_1 + \cdots + X_n$, where the X_i are IID with mean μ then $S_n \approx n\mu$. ie $S_n = n\mu + \text{error}$.
- The Central Limit Theorem refines this, and says that if $\text{Var}[X] = \sigma^2$ then the error is approximately $N(0, n\sigma^2)$. ie. $S_n \approx n\mu + \sigma\sqrt{n}Z$, where $Z \sim N(0,1)$.
- This is the heuristic meaning of the CLT. To make it more precise, we solve for Z and apply the heuristic to calculating probabilities.
- **Thm (CLT):** Let $X_1, X_2, \ldots$ be IID, with mean μ and variance σ^2. Then $P\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z) \to P(Z \le z) = \Phi(z)$ as $n \to \infty$.
- If $\exists$ time at the end of the course, we'll come back to the pf.
- Basically says that for independent sums of many small r.v. we can approximate S_n by a normal r.v. having the same mean and variance as S_n.

Central Limit Theorem

- **Eg:** If the X_k are Bernoulli (ie indicators) then this $\mu = p$ and $\sigma^2 = p(1-p)$. We get $S_n \sim \text{Bin}(n, p)$ and the CLT says exactly the same thing as our Normal approx. to the Binomial.
- **Eg:** In Statistics, one uses frequently that statistics (like $\bar{X}_n = S_n/n$) are asymptotically normal. The CLT is what proves this.
- **Eg:** $Y =$ sum of 20 independent Uniform$([0, 1])$ r.v. Find $P(Y \geq 8)$.

  It is a lot of work to find the exact prob (we did the sum of 2 earlier). A normal approx. is much easier. The uniform mean is $\frac{1}{2}$ [done earlier] and the variance is $\frac{1}{12}$ [done earlier]. So $E[Y] = 10$ and $\text{Var}[Y] = \frac{20}{12} = \frac{5}{3}$. Therefore
  $P(Y \geq 8) = P\left(\frac{Y-10}{\sqrt{5/3}} \geq \frac{8-10}{\sqrt{5/3}}\right) \approx P(Z \geq -1.5492) = 0.9393$
  Note that there is no continuity correction here, as Y already has a density.

Central Limit Theorem

Eg: $\bar{X}$ = the sample mean of 10 indep. r.v. with distribution
$X_k = \begin{cases} 4, & \text{with prob. } \frac{1}{2} \\ -4, & \text{with prob. } \frac{1}{4} \\ 0, & \text{with prob. } \frac{1}{4} \end{cases}$. Find $P(\bar{X} \leq 2)$.

By CLT, S_{10} is approx. normal and therefore so is $\bar{X}$.
$E[X_k] = \frac{4}{2} - \frac{4}{4} + \frac{0}{4} = 1$ and
$\text{Var}[X_k] = (\frac{4^2}{2} + \frac{(-4)^2}{4} + \frac{0^2}{4}) - 1^2 = 11$. So $E[S_{10}] = 10$ and $\text{Var}[S_{10}] = 110$. So $E[\bar{X}] = 1$ and $\text{Var}[\bar{X}] = 1.1$;
$\bar{X}$ is discrete, so we do a continuity correction. The space between neighbouring values of S_{10} is 4, so that between neighbouring values of $\bar{X}$ is 0.4; We split the difference, to apply the normal approx at non-sensitive values. $P(\bar{X} \leq 2) = P(\bar{X} \leq 2.2)$
$= P\left(\frac{\bar{X}-1}{\sqrt{1.1}} \leq \frac{2.2-1}{\sqrt{1.1}}\right) \approx P(Z \leq 1.1442) = 0.8737$

Poisson Distribution

- X has a *Poisson Distribution* with parameter $\lambda > 0$ means that its possible values are $0, 1, 2, \ldots$ (ie any non-negative integer) and $P(X = k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \ldots$
- Why is this a legitimate distribution?
  Need $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = 1$.
  This follows since the Taylor series for e^{λ} is $\sum \lambda^k / k!$
- $E[X] = \lambda$
  [Proof: $E[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda}$. The term with $k = 0$ is 0 so drop it. Then cancel k with $k!$ to give $(k-1)!$ and switch to an index $j = k - 1$.
  ie. $E[X] = \sum_{j=0}^{\infty} \frac{\lambda^{j+1}}{j!} e^{-\lambda} = \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} = \lambda \times 1$]

Poisson Distribution

- Var$[X] = \lambda$.
  [Proof: $E[X^2] = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda}$
  $= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!} e^{-\lambda} + \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} e^{-\lambda}$], because $k^2 = k(k-1) + k$. The 2nd sum is the mean, λ. We drop the 1st 2 terms from the first sum [as they =0], cancel $k(k-1)$ with $k!$ [leaving $(k-2)!$], and change the index to $j = k-2$. This gives $E[X^2] = \lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} + \lambda = \lambda^2 \times 1 + \lambda$. Therefore Var$[X] = (\lambda^2 + \lambda) - \lambda^2 = \lambda$.]
- Why is the Poisson distribution important? It arises when modelling *rare events*.
- An example of this is the **Poisson approximation to the Binomial**.

Poisson approximation to Binomial

- When n is large, and $X \sim \text{Bin}(n, p)$ then $X \approx$ normal. Provided p is NOT ≈ 0 or ≈ 1.
- If $p \approx 0$ the normal approximation is bad, and we turn to a Poisson approximation instead;
  n large, $p = \lambda/n \Rightarrow X \approx \text{Poisson}(\lambda)$.
- More precisely, if $X \sim \text{Bin}(n, p_n)$ and $np_n \to \lambda$ as $n \to \infty$ then $P(X = k) \to \frac{\lambda^k}{k!}e^{-\lambda}$ for every k.
  [Proof: $P(X = k)$
  $= \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1 - p_n)^{n-k}$
  $= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{(np_n)^k}{k!} \cdot \left(1 - \frac{np_n}{n}\right)^n \cdot (1 - p_n)^{-k}$; Let $n \to \infty$.
  The 1st part $= (1)(1 - \frac{1}{n}) \cdots (1 - \frac{k-1}{n}) \to 1$ since each factor does. The 2nd part $\to \frac{\lambda^k}{k!}$; The 3rd part $\to e^{-\lambda}$ (take logs and use l'Hospital's rule); The 4th part $\to 1$. This does it.]

Poisson sums

- If $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$ and X, Y are independent, then $X + Y \sim \text{Bin}(n + m, p)$. Because the number of successes in $n + m$ trials can be broken up as the number in the first n trials, plus the number in the remaining m.
- This suggests that **sums of independent normals are normal**, and also **sums of independent Poisson's are Poisson**. We'll verify the latter.
- If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

  [Proof: By additivity and independence, $P(X_1 + X_2 = n) = \sum_{k=0}^{n} P(X_1 = k, X_2 = n - k)$ $= \sum_{k=0}^{n} P(X_1 = k)P(X_2 = n - k) = \sum_{k=0}^{n} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \cdot \frac{\lambda_2^{n-k} e^{-\lambda_2}}{(n-k)!}$ $= \frac{1}{n!} e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k} = \frac{(\lambda_1+\lambda_2)^n}{n!} e^{-(\lambda_1+\lambda_2)}$, by the binomial theorem. This proves it.]

Poisson scatter

- Let S be a set with a notion of the "size" $|A|$ of its subsets A (eg length, if S is 1-dimensional, area if S is 2-dimensional, etc.). A **Poisson scatter** is a random number N of points chosen from S such that for some λ,
  - No two points can coincide.
  - The number of points in disjoint sets A and B are indep.
  - The mean number of points in any $A \subset S$ is $\lambda|A|$
- In this case, N must be Poisson($\lambda|S|$).
- To see this divide S into n disjoint pieces A_k, each with the same probability p_n of containing a point. By independence, the number of A_k containing points is Bin(n, p_n). Because points can't coincide, this number $\uparrow N$ as $n \to \infty$. By the mean condition, p_n is asymptotically $\lambda|S|/n$. So the result follows from the Poisson limit theorem.
- Is consistent with sums of independent Poisson being Poisson.

Poisson scatter

Eg: Gives a reasonable model of:

- traffic fatalities in Toronto in a month;
- earthquakes in B.C. in a year;

Eg: When a car is painted it has, on average, 1 defect per $10m^2$ (eg bubbles, dirt). Assuming that defects occur independently of each other, what is the probability that a car with area $4m^2$ has at least 2 defects

- The Poisson scatter properties hold. So the total number of defects N has a Poisson distribution. We're given that the average number per m^2 is $\lambda = \frac{1}{10} = 0.1$; So $E[N] = 4\lambda = 0.4$; Therefore $P(N \geq 2) = 1 - P(N = 0) - P(N = 1)$ $= 1 - e^{-0.4} - 0.4 \times e^{-0.4} = 0.0616$

Geometric distribution

Under construction

- Geometric Distribution on $\{1, 2, 3, \ldots\}$
- Geometric Distribution on $\{0, 1, 2, \ldots\}$
- Models time to 1st success (or 1st failure)
- Mean and Variance
- **Eg:** Flip a coin, wait for 1st Head.
- **Eg:** Craps (dice game)

Exponential distribution

Under construction

- Exponential distribution: density
- Used in actuarial science (lifetimes), queueing theory (service times), reliability theory (failure times), etc.
- Survival probabilities, $\lambda =$ exponential decay rate.
- mean and variance
- memoryless property
- constant hazard rate
- [described more general hazard rates, ie ageing, but you're not responsible for this]

Exponential distribution

Under construction

- **Eg:** A person age 65 has an expected remaining lifetime of 20 years (ie to age 85). What's the probability they live to age at least 90?
- **Eg:** We measure radioactive decay from a lump of uranium ore, and find that it takes on average 10 minutes till the first decay. What's the probability of a decay in the first 5 minutes?

Exponential and Poisson

- The Poisson and Exponential are related. Let arrival times be distributed on $[0, \infty)$ according to a Poisson scatter with rate λ. Let N_t be the number of arrivals before time t. Then N_t is Poisson(λt), and from that we can get that the time T_1 of the first arrival is Exponential(λ).
- To see this, observe that $P(T_1 > t) = P(\text{no arrivals in } [0, t]) = P(N_t = 0) = e^{-\lambda t}$. From this we get that T_1 has an exponential cdf, and so an exponential density.
- More generally, if T_k is the time between the $k-1$st and kth arrivals, then the T_k are independent Exponential(λ).
- Let S_k be the time of the kth arrival, so $S_1 = T_1$, $S_2 = T_1 + T_2$, $S_3 = T_1 + T_2 + T_3$, etc. We can work out the density for S_k. It gives us an example of what's called a *Gamma* distribution. So sums of independent exponentials (with the same λ) are Gamma.

Gamma

- For example, $P(S_2 > t) = P(\text{at most 1 arrival in } [0, t])$ $= P(N_t = 0) + P(N_t = 1) = (1 + \lambda t)e^{-\lambda t}$. So the cdf of S_2 is $F(s) = \begin{cases} 0, & t < 0 \\ 1 - (1 + \lambda t)e^{-\lambda t}, & t \geq 0. \end{cases}$
- This gives the density $f(s) = \begin{cases} 0, & t < 0 \\ \lambda^2 t e^{-\lambda t}, & t > 0. \end{cases}$
- In general, a Gamma density has the form
  $f(s) = \begin{cases} 0, & t < 0 \\ C(\alpha, \beta) t^{\alpha - 1} e^{-t/\beta}, & t > 0. \end{cases}$
  for parameters α and β. Here C is a constant that makes this integrate to 1.
- The sum of k independent exponentials is then Gamma, with $\alpha = k$ and $\beta = 1/\lambda$.

Negative Binomial

- Another example of a Gamma distribution is the Chi-squared distribution, from statistics. Now $\alpha = \frac{1}{2}$. [To see this, do exercise 10b of §4.4]
- In the discrete setting one can do similar things: Carry out independent trials and let N_k be the time of the kth success. One can calculate its distribution, called the *Negative binomial*.
- One can show that N_k is also the sum of k independent Geometric r.v.
- $P(N_k = n) = P(n\text{th trial is S, \& } k-1 \text{ S's in 1st } n-1 \text{ trials})$
$= \binom{n-1}{k-1}(1-p)^{n-k}p^k$, $n = k, k+1, \ldots$

[Note: You are not responsible for the Gamma or Negative Binomial distributions]

Discrete joint distributions

The joint distribution of a pair of r.v. X, Y is a table of values $P(X = x, Y = y)$. Use it to:

- Calculate expectations
  $E[g(X, Y)] = \sum_{x,y} g(x, y)P(X = x, Y = y)$.
  [Proof. $g(X, Y) = \sum_{x,y} g(x, y)1_{\{X=x,Y=y\}}$. To see this, substitute ω. The LHS is $g(X(\omega), Y(\omega))$. All terms on the RHS $= 0$ except the one with $x = X(\omega)$ and $y = Y(\omega)$. And that gives $g(x, y) = g(X(\omega), Y(\omega))$. Now take expectations.]
- Verify independence.
  ie. is $P(X = x, Y = y) = P(X = x)P(Y = y)$?
- Calculate *marginal distributions* $P(X = x)$ and $P(Y = y)$.
  ie. sum over rows or columns, to get
  $P(X = x) = \sum_y P(X = x, Y = y)$ and
  $P(Y = y) = \sum_x P(X = x, Y = y)$.

Discrete joint distributions

- Calculate *conditional distributions* $P(Y = y \mid X = x) = \frac{P(X=x, Y=y)}{P(X=x)}$.
- Find the *covariance* $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$ between X and Y.
- Find the *correlation* $\rho(X, Y)$ between X and Y. That is, find $\rho(X, Y) = \frac{\text{Cov}(X,Y)}{\text{SD}[X] \cdot \text{SD}[Y]}$.
- We'll see that $-1 \leq \rho \leq 1$, and that ρ measures the extent to which there is a *linear relationship* between X and Y: $\rho = 0$ means they're *uncorrelated*; there is no linear relationship between them. $\rho = 1$ means they're perfectly positively correlated; there is a perfect linear relationship between them (with positive slope). $\rho = -1$ means they're perfectly negatively correlated; there is a perfect linear relationship between them (with negative slope). Other ρ's reflect a partial linear relationship with varying degrees of strength.

Covariance

Properties of covariance:

- $\text{Var}[X] = \text{Cov}(X, X)$. [Pf: By definition]
- X, Y independent $\Rightarrow \text{Cov}(X, Y) = 0 \Rightarrow \rho(X, Y) = 0$.
  [Pf: $= E[X - E[X]] \cdot E[Y - E[Y]]$ by indep. This $= 0 \times 0$.]
- $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}(X, Y)$. [Pf:
  $= E[((X+Y) - E[X+Y])^2] = E[((X - E[X]) + (Y - E[Y]))^2]$.
  Now expand the square and match up terms.]
  This is consistent with our earlier observation that
  independence $\Rightarrow$ Var of sum $=$ sum of Var.
- $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$
  [Pf: Expand, so $= E[XY - XE[Y] - YE[X] + E[X]E[Y]] =$
  $E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]$, and just cancel
  the last 2 terms.]

Correlation

Properties of correlation:

- $-1 \leq \rho(X, Y) \leq 1$.
- $\rho = 1 \Rightarrow$ the values of X and Y always like on an upward sloping line. [They are perfectly positively correlated]
- $\rho = -1 \Rightarrow$ the values of X and Y always like on a downward sloping line. [They are perfectly negatively correlated]

[Pf: Let μ_X, σ_X, μ_Y, σ_Y be the mean and S.D. of X and Y. Then $0 \leq E\Big[(\frac{X-\mu_X}{\sigma_X} - \frac{Y-\mu_Y}{\sigma_Y})^2\Big] = \frac{\text{Var}(X)}{\sigma_X^2} + \frac{\text{Var}(Y)}{\sigma_Y^2} - 2\frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$ $= 1 + 1 - 2\rho(X, Y)$. So $2\rho \leq 2$, which shows that $\rho(X, Y) \leq 1$. The only way we could have $\rho = 1$ is if the above expectation $= 0$, which implies that $\frac{X-\mu_X}{\sigma_X} = \frac{Y-\mu_Y}{\sigma_Y}$, a linear relationship. The same argument but with $E\Big[(\frac{X-\mu_X}{\sigma_X} + \frac{Y-\mu_Y}{\sigma_Y})^2\Big]$ shows $\rho \geq -1$.]

Example

Eg: Suppose $P(X = x, Y = y)$ is

	-1	0	1	x
1	0	$\frac{1}{7}$	$\frac{1}{7}$	
0	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	
-1	$\frac{1}{7}$	$\frac{1}{7}$	0	
y				

Then:

- X and Y aren't independent, since $P(X = -1, Y = 1) = 0 \neq P(X = -1)P(Y = 1)$.
- $E[XY] = (-1) \times (-1) \times \frac{1}{7} + (-1) \times 0 \times \frac{1}{7} + (-1) \times 1 \times 0$ $+0 \times (-1) \times \frac{1}{7} + 0 \times 0 \times \frac{1}{7} + 0 \times 1 \times \frac{1}{7}$ $+1 \times (-1) \times 0 + 1 \times 0 \times \frac{1}{7} + 1 \times 1 \times \frac{1}{7} = \frac{2}{7}$.
- Adding up each column, we get that the marginal distribution of X is

x	-1	0	1
$P(X = x)$	$\frac{2}{7}$	$\frac{3}{7}$	$\frac{2}{7}$

Example (cont'd)

- Adding up each row gives the marginal distribution for Y, which is the same as that of X.
- Therefore $E[X] = (-1) \times \frac{2}{7} + 0 \times \frac{3}{7} + 1 \times \frac{2}{7} = 0$. Likewise $E[Y] = 0$.
- So $\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{2}{7} - 0 = \frac{2}{7}$. The fact that this $\neq 0$ also tells us that X and Y aren't independent.
- $E[X^2] = \frac{2}{7} + 0 + \frac{2}{7} = \frac{4}{7}$ and $E[X] = 0$, so $\text{Var}[X] = \frac{4}{7}$. LIkewise for Y.
- So $\rho(X, Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{2/7}{4/7} = \frac{1}{2}$.
- Therefore the r.v. X and Y are positively correlated, but not perfectly so.

Example (cont'd)

- If instead, $P(X = x, Y = y)$ was

	-1	0	1	x
1	0	0	$\frac{1}{3}$	
0	0	$\frac{1}{3}$	0	
-1	$\frac{1}{3}$	0	0	
y				

Then the same calculations show:

- $E[XY] = \frac{2}{3}$, $\text{Var}[X] = \frac{2}{3} = \text{Var}[Y]$, so
  $\text{Cov}(X, Y) = \frac{2/3}{\sqrt{(2/3)(2/3)}} = 1.$
  In other words, now X and Y are perfectly correlated.
- In fact, in this case $Y = X$ so there is indeed a linear relation between them.
- More generally, if $Y = aX + b$ then $\rho = 1$ when $a > 0$ and $\rho = -1$ when $a < 0$.