Math 425
Intro to Probability
Lecture 28

Kenneth Harris
kaharri@umich.edu

Department of Mathematics
University of Michigan

March 27, 2009

Theorem: convolutions and sums of r.v.'s

Theorem

Let X and Y be independent continuous random variables with density $f_X(x)$ and $f_Y(y)$.

The sum $X + Y$ is a continuous random variable with density $f_{X+Y} = f_X * f_Y$. That is

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\, dy$$

Theorem

Theorem (Normal Random Variables)

Let $X_1, X_2, \ldots, X_n$ be independent normal random variables with respective parameters μ_i, σ_i^2, $i = 1, \ldots, n$.

Then the sum S_n of these random variables is a normal random variable with pararmeters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$.

☞ See Ross Proposition 6.3.2, page 283.

Proof of a special case

☞ When X is a normally distributed random variable with mean μ and variance σ^2, the density is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Ross proves the most general case in Proposition 3.2.

☞ I will prove the case of two standard normal random variables, X and Y, whose density is

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

with mean $\mu = 0$ and variance $\sigma^2 = 1$.

Proof of a special case

Then the density of $X+Y$ is

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(a-y)^2/2} e^{-y^2/2}\,dy \\ &= \frac{1}{2\pi} e^{-a^2/2}\int_{-\infty}^{\infty} e^{-(y^2-ay)}\,dy \\ &\overset{1}{=} \frac{1}{2\pi} e^{-a^2/4}\int_{-\infty}^{\infty} e^{-(y-(a/2))^2}\,dy \\ &\overset{2}{=} \frac{1}{2\pi} e^{-a^2/4}\sqrt{\pi}\left[\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-(y-(a/2))^2}\,dy\right] \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-a^2/4} \qquad (\text{normal: } \mu = 0, \sigma^2 = 2) \end{aligned}$$

(1): Complete the square: $-(y^2 - ay) = -(y^2 - ay + (a/2)^2) + (a/2)^2$
(2): $[\ldots] = 1$: normal with $\mu = \frac{a}{2}$ and $\sigma^2 = \frac{1}{2}$.

Proof of a special case

☞ When X and Y are normally distributed random variables with mean $\mu = 0$ and $\sigma^2 = 1$, their sum $X+Y$ has density

$$f_{X+Y}(a) = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-a^2/4}$$

and is a normally distributed random variable with $\mu = 0$ and $\sigma^2 = 2$.

Proof of a special case

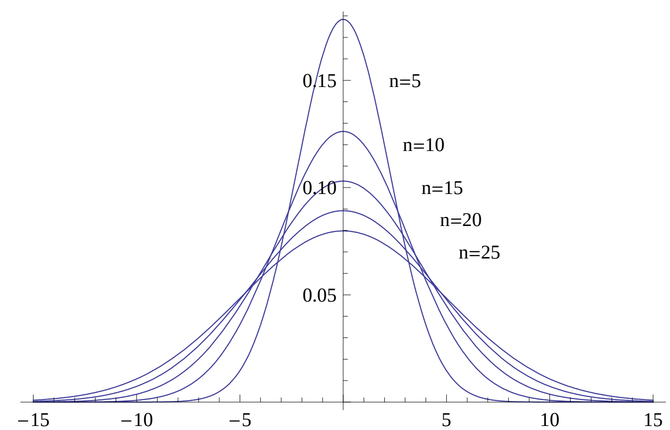
☞ Let $X_1, X_2, \ldots, X_n$ be independent and distributed normally with $\mu = 0$ and $\sigma^2 = 1$.
The sum $S_n = X_1 + X_2 + \ldots + X_n$ has a density function

$$f_{S_n}(a) = \frac{1}{\sqrt{2\pi n}} e^{-a^2/(2n)}$$

☞ The sum S_n is normally distributed with $\mu = 0$ and $\sigma^2 = n$.
This is the sum of the means and variances of the n standardly distributed random variables.

Example – continued

☞ Sum of $n = 5, 10, 15, 20, 25$ independent and normally distributed random variables with $\mu = 0$ and $\sigma^2 = 1$.



Example

Example. Let X and Y be independent random variables uniformly distributed on the interval $[-1, 1]$. What is the density of their sum $X + Y$?

☞ The common density is

$$f_X(x) = f_Y(x) = \begin{cases} \frac{1}{2} & \text{if } -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

The density of the sum is the convolution $f_{X+Y} = f_X * f_Y$,

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy$$

and since the sum is bounded by $-2 \le X + Y \le 2$,

$$f_{X+Y}(z) = 0 \qquad \text{unless } -2 \le z \le 2.$$

Example – continued

☞ Let $-2 \le z \le 2$. Since $-1 \le y \le 1$

$$f_{X+Y}(z) = \frac{1}{2} \int_{-1}^{1} f_X(z-y)\, dy.$$

But, $-1 \le z - y \le 1$ as well, or equivalently

$$z - 1 \le y \le z + 1.$$

☞ So, for $-2 \le z \le 2$, $f_{X+Y}(z) = 0$ unless

$$\max\{z-1, -1\} \le y \le \min\{z+1, 1\}.$$

Example – continued

☞ If $-2 \le z \le 0$, we have $(z - 1 < -1)$

$$f_{X+Y}(z) = \frac{1}{4} \int_{-1}^{z+1} dy = \frac{z+2}{4}.$$

☞ If $0 \le z \le 2$, we have $(z + 1 > 1)$

$$f_{X+Y}(z) = \frac{1}{4} \int_{z-1}^{1} dy = \frac{2-z}{4}.$$

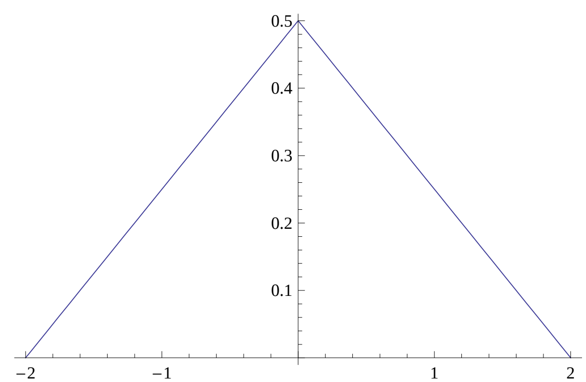
☞ The density of $X + Y$ is

$$f_{X+Y}(z) = \begin{cases} \frac{z+2}{4} & \text{if } -2 \le z \le 0, \\ \frac{2-z}{4} & \text{if } 0 < z \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Example – continued

☞ The density of $X + Y$ is

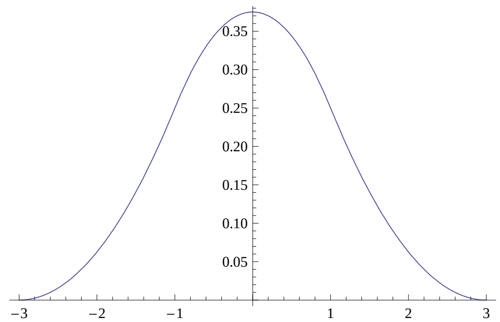
$$f_{X+Y}(z) = \begin{cases} \frac{z+2}{4} & \text{if } -2 \le z < 0, \\ \frac{2-z}{4} & \text{if } 0 \le z \le 2 \\ 0 & \text{otherwise.} \end{cases}$$



Example – continued

☞ Let X, Y, Z be independent and uniformly distributed on $[-1, 1]$. Then

$$f_{X+Y+Z}(w) = \begin{cases} \frac{(w+3)^2}{16} & \text{if } -3 \le z < -1, \\ \frac{3-w^2}{8} & \text{if } -1 \le z < -1, \\ \frac{(3-w)^2}{16} & \text{if } 1 \le z < 3, \\ 0 & \text{otherwise.} \end{cases}$$



Example – continued

☞ Let X be uniformly distributed on $[-1, 1]$. Then

$$E[X] = 0 \qquad Var(X) = \frac{1}{3}.$$

☞ Let X, Y, Z be independent and uniformly distributed on $[-1, 1]$. Then

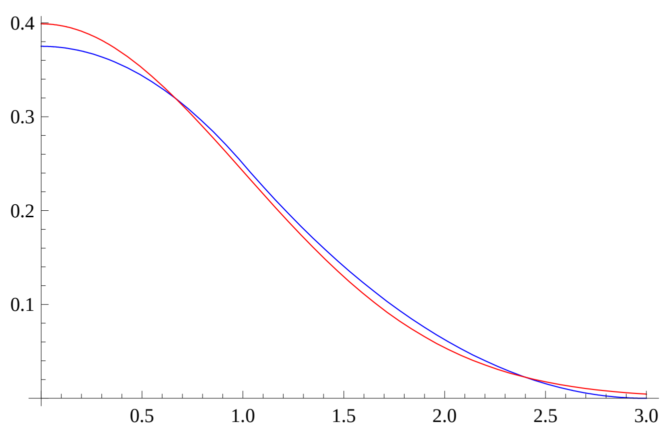
$$E[X+Y+Z] = 3 \cdot E[X] = 0 \qquad Var(X+Y+Z) = 3 \cdot Var(X) = 1.$$

We will prove this in Chapter 7. You can verify by using the density on the previous slide ☺

Example – continued

☞ Let X, Y, Z be independent and uniformly distributed on $[-1, 1]$.

Comparison of the density function f_{X+Y+Z} and the standard normal density ($\mu = 0, \sigma^2 = 1$) on the interval $[0, 3]$.



Example – continued

☞ Sums of uniform random variables are not so nice to compute ☹.

The sum S_n of n independent and identically distributed uniform random variables on the interval $[0, 1]$ has “nice" form

$$f_{S_n}(x) = \begin{cases} \frac{1}{(n-1)!} \sum_{j=0}^{x} (-1)^j \binom{n}{j} (x-j)^{n-1} & \text{if } 0 < x < n, \\ 0 & \text{otherwise.} \end{cases}$$

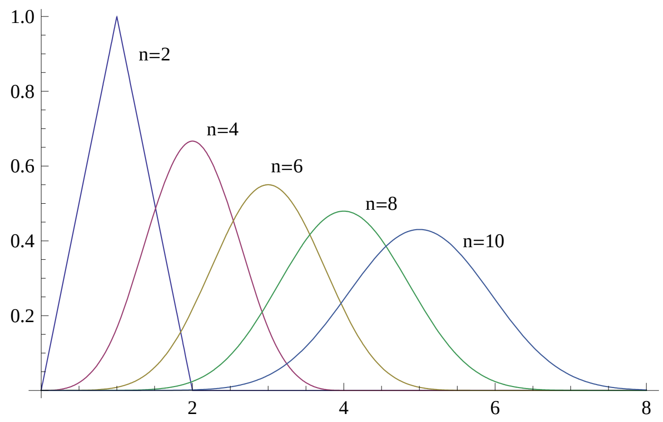
☞ When $n = 2$ (see Ross, Example 6.3a),

$$f_{S_2}(x) = \begin{cases} x & \text{if } 0 \le x \le 1, \\ 2 - x & \text{if } 1 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Example – continued

☞ The density functions for the sums of $n = 2, 4, 6, 8, 10$ independent and uniformly distributed random variables on $[0, 1]$.

They are looking more "normal" ☺



Central Limit Theorem

☞ The uniform distribution provides more evidence for the Central Limit Theorem:

If you average enough distributions together, even if they are not normal, in the limit their average is normal.

☞ The Central Limit Theorem is a limit theorem:

- The distributions converge to the normal distribution in the limit, and this limit may be slow depending on what is being average.

(The uniform distribution converges quickly, for example; the exponential distribution not so quickly.)

Fuzzy Central Limit Theorem

☞ The Fuzzy Central Limit Theorem "explains" why we expect so many quantitative properties to be normally distributed:

Data that are influenced by many small and unrelated random effects are approximately normally distributed.

☞ For example, a monkey is throwing a dart at a dart board and we measure the distance from where the dart lands to the bull's eye. Many small factors can influence can effect the position of each dart:

✏ monkey is thinking of a bunch of bananas, drafts, drifting smoke blurs vision, number of beers monkey has had, someone yells at the next dart board, dart slightly imbalanced, etc.

☞ If we measure locations over many games, what is the distance from the distribution of distances from the bull's eye?

Example

Example

Suppose a monkey throws a dart at a dartboard, the xy-plane, whose bulls-eye is the origin. If the monkey is a decent dart thrower, we might expect the x and y coordinates of his throws to be independent and normally distributed with $\mu = 0$ and $\sigma^2 = 1$.

This problem is more serious than it appears. The problem arises in physics when it is assumed that a moving particle in $\mathbb{R}^3$ has velocity components that are mutually independent and normally distributed, and we want to find the distribution of the speed of the particle. This is called the Maxwell density.

We will stick to $\mathbb{R}^2$ and monkeys. Let X and Y be independent and normally distributed. We will compute the density of the distance R from the origin, where

$$R^2 = X^2 + Y^2.$$

Example

☞ The density of X and Y are

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The density of X^2 is computed in Ross (Section 5.7):

$$\begin{aligned} f_{X^2}(r) &= \begin{cases} \frac{1}{2\sqrt{r}}\left(f_X(\sqrt{r}) + f_X(-\sqrt{r})\right) & \text{if } r > 0, \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{1}{\sqrt{2\pi r}} e^{-r/2} & \text{if } r > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

☞ This is the gamma density with $(\alpha = \frac{1}{2}, \lambda = \frac{1}{2})$:

$$f(r) = \frac{\lambda e^{-\lambda r}(\lambda r)^{\alpha-1}}{\Gamma(\alpha)} \qquad r > 0,$$

since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. (See Lecture 24, slide 33).

Example – continued

☞ Each of X^2 and Y^2 are independent, let a, b be real numbers

$$\begin{aligned} \mathbf{P}\left\{X^2 \le a, Y^2 \le b\right\} &= \mathbf{P}\left\{-a \le X \le a, -b \le Y \le b\right\} \\ &= \mathbf{P}\left\{-a \le X \le a\right\} \cdot \mathbf{P}\left\{-b \le Y \le b\right\} \\ &= \mathbf{P}\left\{X^2 \le a\right\} \cdot \mathbf{P}\left\{Y^2 \le b\right\} \end{aligned}$$

and gamma distributed with $(\alpha = \frac{1}{2}, \lambda = \frac{1}{2})$.

Example – continued

☞ Each of X^2 and Y^2 are independent and gamma distributed with $(\alpha = \frac{1}{2}, \lambda = \frac{1}{2})$.

By Proposition 6.3.1 (Ross, p. 281 and Lecture 27, slide 29) R^2 is gamma distributed with $(\alpha = 1, \lambda = \frac{1}{2})$

$$\begin{aligned} f_{R^2}(r) &= \int_{-\infty}^{\infty} f_{X^2}(r-s) \cdot f_{Y^2}(s)\, ds \\ &= \frac{1}{4\pi} \int_{-\infty}^{\infty} e^{-(r-s)/2} \frac{(r-s)^{-1/2}}{2} \cdot e^{-s/2} \frac{s^{-1/2}}{2}\, ds \\ &= \begin{cases} \frac{1}{2} e^{-r/2} & \text{if } r \ge 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Rayleigh density

☞ The density of the random variable R can be obtained from that of R^2 using Theorem 5.7.1 (Ross, p. 243).

Let $g(x) = \sqrt{x}$, so g is increasing and $g^{-1}(y) = x^2$.

$$\begin{aligned} f_R(r) &= \begin{cases} f_{R^2}(g^{-1}(r)) \cdot \frac{d}{dr} g^{-1}(r) & \text{if } r > 0 \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{1}{2} e^{-r^2/2} \cdot 2r = r e^{-r^2/2} & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

☞ This is the Rayleigh distribution, and is the density function for the distance of the point P from the origin when the coordinates of P are normally distributed with $\mu = 0, \sigma^2 = 1$.

Chi-Squared Density

☞ More generally, if $Z_1, Z_2, \ldots, Z_n$ are independent and standard normal random variables, then $X = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ has a gamma distribution with $(\alpha = \frac{n}{2}, \lambda = \frac{1}{2})$:

$$f_X(r) = \frac{e^{-r^2/2}\, r^{n/2-1}}{2^{n/2}\,\Gamma(\frac{n}{2})}$$

where (see Lecture 27, slide 29)

$$\Gamma(\frac{n}{2}) = \begin{cases} [(n/2)-1]! & \text{when } n \text{ is even,} \\ (\frac{n-2}{2})\cdot(\frac{n-4}{2})\cdots\frac{3}{2}\frac{1}{2}\sqrt{\pi} & \text{when } n \text{ is odd} \end{cases}$$

☞ This is the density function for the chi-squared distribution (written χ^2) with n degrees of freedom. It is the distribution of the square of the distance of a point P from the origin in n-dimensional space.

Example – Chi-Squared Distribution

☞ An important use of the chi-squared distribution is in comparing experimental data with a theoretical discrete distribution, to determine whether the data supports the model.

☞ Let X be a finite discrete random variable on ν outcomes with mass function $m(i)$, our theoretical distribution. We have a sample size n and have observed o_i outcomes of type i.

Then for moderately large n,

$$V = \sum_{i=1}^{\nu} \frac{(o_i - n\cdot m(i))^2}{n\cdot m(i)} \qquad \frac{\text{square of distance}}{\textit{weightfactor}}$$

is approximately chi-squared distributed with $\nu - 1$ degrees of freedom.

This is NOT obvious. Intuitively, V measures the average deviation of the measured frequency from the theoretical frequency.

Example – Chi-Squared Distribution

Example

Suppose we have a given single die. We wish to test the hypothesis that the die is fair. The theoretical distribution is the uniform distribution on the integers $\{1, 2, \ldots, 6\}$.

We roll the die 60 times and observe the following frequency of outcomes:

Outcome	Frequency	Outcome	Frequency
1	15	4	5
2	8	5	7
3	7	6	18

Is the die fair?

Example – continued

☞ Let o_i denote the actual number of data points of type i, for $1 \le i \le 6$. The theoretical distribution ($p(i) = \frac{1}{6}$) predicts each of these values should be $60 \cdot \frac{1}{6} = 10$. The expression

$$V = \sum_{i=1}^{6} \frac{(o_i - 10)^2}{10}$$

is approximately the chi-squared distribution with 5 degrees of freedom.

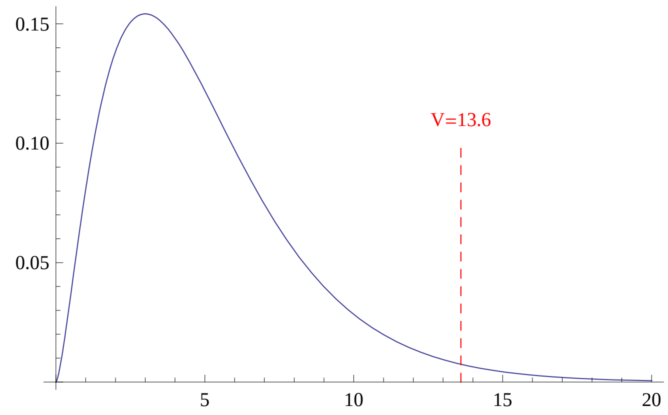
☞ For this example, $V = 13.6$. But

$$\chi^2(13.6) \approx 0.0074,$$

so we can be better than 99% certain that a fair die would NOT have this observed frequency.

Example – continued

☞ Chi-squared distribution with 5 degrees of freedom.



Example

Example

Assume that the service time for a customer at a bank is exponentially distributed with a mean service time of 2 minutes. Let X be the total service time for 10 customers.

What is the probability that $X > 22$?

☞ We are assuming the times helping customers are independent and exponentially distributed with $\lambda = \frac{1}{2}$.

Recall, the expected time is $\frac{1}{\lambda}$.

☞ X (the waiting time of 10 events) is gamma distributed with $(\alpha = 10, \lambda = \frac{1}{2})$, so has the density function

$$f_X(t) = \frac{1}{2} e^{-\frac{1}{2}t} \frac{(\frac{1}{2}t)^9}{9!} \qquad \text{when } t > 0.$$

Example – continued

☞ You could compute the integral

$$\mathbf{P}\{X > 22\} = \frac{1}{2 \cdot 9!} \int_{22}^{\infty} e^{-\frac{1}{2}t} (\frac{1}{2}t)^9 \, dt.$$

There is an easier way.

☞ Let $N(t)$ be the Poisson random variable (with $\mu = 22 \cdot \frac{1}{2} = 11$) which counts the number of customers helped in t minutes. So,

$$\begin{aligned} \mathbf{P}\{X > 22\} &= \mathbf{P}\{N(22) \leq 9\} \\ &= \sum_{i=0}^{9} e^{-\mu} \frac{\mu^j}{j!} \\ &= \sum_{i=0}^{9} e^{-11} \frac{11^j}{j!} \\ &\approx 0.3405 \end{aligned}$$

Example

Example

The gross weekly sales at a restaurant is a normal random variable with mean \$2200 and standard deviation \$230.

What is the probability that the total salses over the next two weeks exceed \$5000?

☞ Let X_1 and X_2 be the random variables which denote the sales in each week. Then $X_1 + X_2$ is also a normal random variable with mean \$4400 and standard deviation $\sqrt{2 \cdot 230^2} \approx 325$.

$$\begin{aligned} \mathbf{P}\{X_1 + X_2 > 5000\} &= \mathbf{P}\{Z > \frac{600}{325}\} \\ &\approx 1 - \Phi(1.85) = 0.0322 \end{aligned}$$