# Continuous Probability Models & ...

... Operations on Random Variables
Week 6

Prof. Brand

ADM2303

February 12, 2012



## Administrative Issues

1. Assignment-2
   - Due Feb 20th
   - Relative weight apportioned to Part-I & II
2. Quiz (Sunday Mar ... see doc-depot)
3. Reading (Ch9.10-9.11).



## Last Time

1. Wrapping up Discrete probability models (PMFs)
2. Continuous probability models (PDF's)
   - Uniform (Ch9.9)
   - Normal (Ch9.10)
   - Summary of PDF's
3. Contrast between Discrete and Continuous



## This Week

1. Normal probability model
   - Area questions
   - ... as approximation of binomial
2. Exponential probability model
   - Model and area questions
   - *Link with Poisson model*
3. Operations on RV's (correlated)
   - Bivariate Probability models
   - covariance
   - correlation
   - Special case of Normal RV's
4. The inverse problem



## Normal Probability Density Function (PDF)

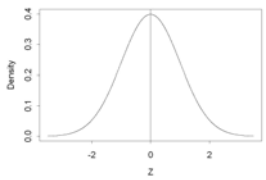
A normal probability density function has two parameters: $\mu_x$ and $\sigma_x$.

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

bounds $-\infty < x < \infty$

mean $\mu_x = \mu_x$

variance $\sigma_x^2 = \sigma_x^2$






## Gauging Area: The 68/95/99 Rule

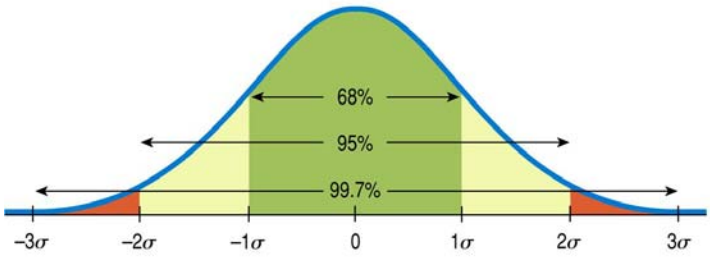
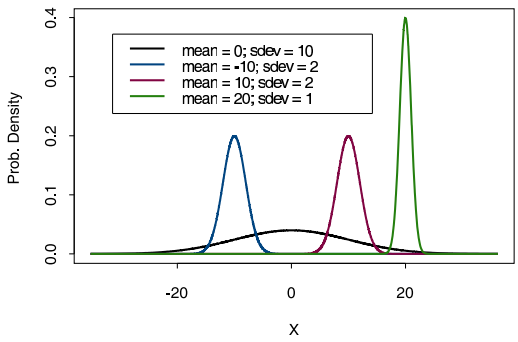



Figure: Reproduced from Sharpe et al, 2009

## Family of Normal PDFs






## Standardized Normal

- Since the number of possible $\mu$'s and $\sigma$'s (each defining a unique normal model) is unlimited, there are an unlimited number of unique normal PDFs,
- We could tabulate "area's" for a single normal model for various values of $X$, but we could not do so for all models,
- Why not tabulate "area's" for a standard normal PDF, and then use the standard results to address the general case,
- Thus, the motivation behind the **"Standard Normal Distribution"**



## Z-Score

$$z = \frac{(x - \mu_x)}{\sigma_x}$$

- The numerator represents how different $x$ is from its expected value ($\mu_x$),
- Dividing by $\sigma_x$ scales that difference in units of the standard deviation,
- Thus, a z-score gives the number of standard deviations that a value x, is above or below the mean.



## Calculating Z-Score

Suppose $\mu = 20$ cars, and $\sigma = 5$ cars, and further that you are interested in the probability that $X > 15$ cars.
The first steps

1. Sketch normal, and shade in area of interest
2. Compute your Z-score

$$\begin{aligned} z &= \frac{(x - \mu_x)}{\sigma_x} \\ &= \frac{(15 - 20)\text{ cars}}{5\text{ cars}} = -1 \end{aligned}$$



## Standardized Normal

- If X is distributed with a mean $\mu_x$ and standard deviation of $\sigma_x$ then the z-score will also be normally distributed with a mean of ... and a standard deviation of ....
- Since any normal distribution can be standardized (converted to a standard normal), tables have been developed for the standard normal. The tables help us with **area** questions.
- See Appendix of your text (Appendix C in Sharpe et al).

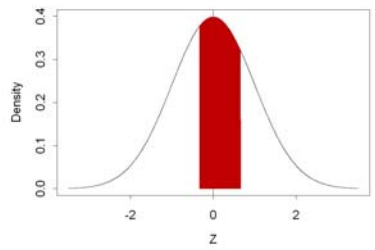


## Chapter 9 Examples

See page 256-260 in Sharpe.
(Note, excellent Plan/Do/Report example [Cereal Company] on page 257- in Sharpe).

## Example

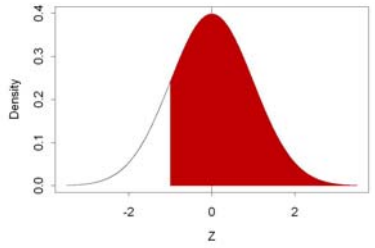
1 If X is a normal random variable with parameters $\mu_x = 3$ and $\sigma_x = 3$, find:
   (a) $P[2 < X < 5]$,






## Same model, different area

1 X is a normal RV with parameters $\mu_x = 3$ and $\sigma_x = 3$, find:
   (b) $P[X > 0]$.






## Blues Fest Again: Demand

- A restaurant owner is considering opening up a booth at the Blues festival. She estimates the probability that a concert-goer would purchase her chicken sandwiches: $p = 0.05$;
- Last year the Blues festival attendance was 400000 people;
- What is the probability that her booth will have greater than 20300 sales?

To begin solution note:

- $np = 20000 > 10$
- $n(1-p) > 10$



## Approximating the Binomial

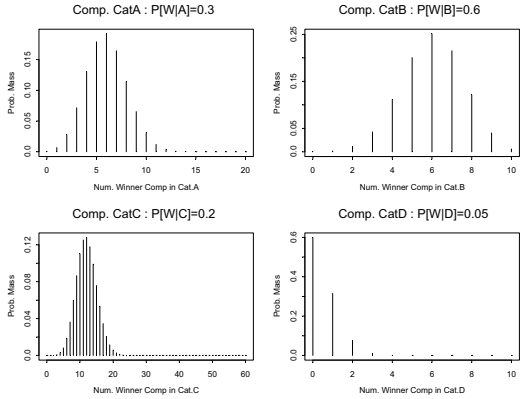
**Normal** If expected number of successes and failures sufficiently high ($np > 10$ and $n(1-p) > 10$) then normal provides adequate approximation.

**Poisson** If expected number of successes/failures is small ($np < 10$) the the Poisson model can serve as an approximation.



## How Good of An Approximation

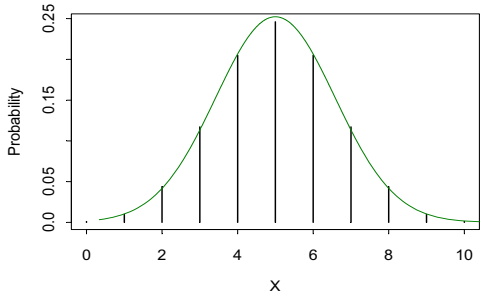
An old Example. Binomial probability models.





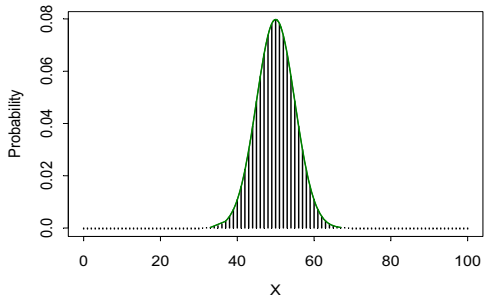

## How Good of An Approximation

$p = 0.5, n = 10$



## How Good of An Approximation

$p = 0.5, n = 100$






## Blues Fest Again: Demand

mean ($\mu_x$) $np = 20000$ customers

variance ($\sigma_x^2$) $np(1-p) = 19000$ customers$^2$

sd ($\sigma_x$) $\sqrt{np(1-p)} = 138$ customers

Calculate z-score

$$\begin{aligned} z &= \frac{(x-\mu_x)}{\sigma_x} \\ &= \frac{(20300-20000)}{138} \\ &\approx 2 \end{aligned}$$

Recall 68/**95**/99 rule. It implies that the mean $\pm 2\sigma$ encloses 95% probability, leaving 5% to be split between left and right tails.



## Blues Fest Again: Demand

Exact (table-based solution)

$$\begin{aligned} z &= \frac{(20300-20000)}{138} \\ &\approx 2.174 \end{aligned}$$

$$\begin{aligned} \Pr[X < 20300] &= \Pr[Z < 2.174] \\ F(Z = 2.174) &= \Pr[Z < 2.174] \\ &= 0.9850 \end{aligned}$$

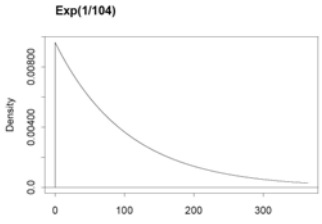
Thus the probability of *exceeding* 20300 sales is 0.015 (small!).



## Exponential PDF

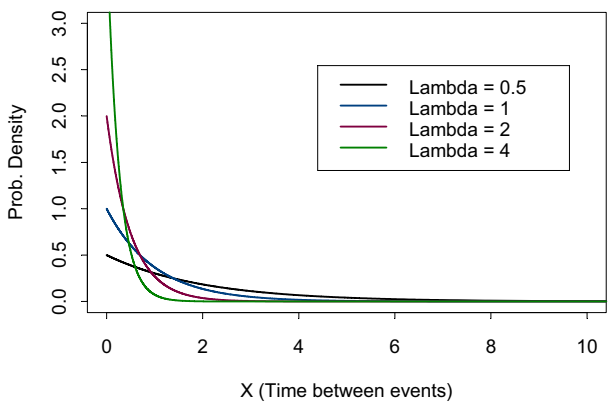
The exponential PDF has a single parameter denoted $\lambda$ (where $\lambda > 0$) and can be expressed as:

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$






## Family of Exponential PDFs





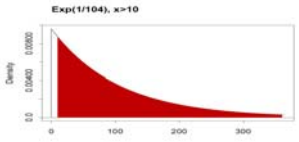

## Summary Measures

mean $\mu_x = \frac{1}{\lambda}$

variance $\sigma_x^2 = \frac{1}{\lambda^2}$

Coefficient of variation $\nu_x = 1$ (confirm for yourself)

## Calculating Area




Handy formula for *cumulative density function*. Helps address questions about the probability that $X \leq a$.

$$\begin{aligned} F(a) &= P[X \leq a] \\ &= \int_0^a \lambda \exp(-\lambda x) dx \\ &= 1 - \exp(-\lambda a) \text{ where } a \geq 0 \end{aligned}$$



## When is Exponential PDF Applicable?

The exponential PDF often arises in practice, as being the distribution of the amount of time until some specific event occurs[1]. For example time until the next:

- Tsunami, earthquake, or the next 'wrong-number' telephone call, or
- until the next car to join a que, or
- the next reaction between two molecules in a room, or
- until the next match between a salesperson and a client

The exponential PDF has a link to the Poisson PMF (we saw the Poisson PMF previously).

[1] Note the parallel between the geometric model (number of trials until next success) and the Exponential distribution (time until the next success).



## Phone Call

Suppose that the length of a phone call in minutes ($X$) is an exponential RV with parameter $\lambda = 1/104$ (1/minutes).

(a) What is the variance of X?

If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait:

(b) more than 10 minutes;

(c) between 10 and 20 minutes.



## Phone Call: Solution (a)

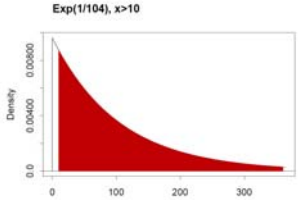
$$Var[X] = 1/\lambda^2 = (104)^2 \text{ minutes}^2$$



## Phone Call: Solution (b)

$$\begin{aligned} P[X > 10] &= 1 - P[X < 10] = 1 - (1 - \exp(-\lambda 10)) \\ &= \exp\left(-\frac{10}{104}\right) = 0.91 \end{aligned}$$

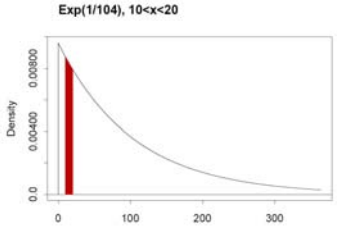
So that you are very likely to have to wait more than 10 minutes! Not surprising.






## Phone Call: Solution (c)

$$\begin{aligned} P[10 < X < 20] &= P[X < 20] - P[X < 10] \\ &= \left(1 - \exp\left(-\frac{20}{104}\right)\right) - \left(1 - \exp\left(-\frac{10}{104}\right)\right) \end{aligned}$$



## What if RVs are Not Independent?

- Rule for expected-value remains **unchanged**.
- Calculation of variance must be **altered**.
- Consider $X$ and $Y$ as dependent RVs

$$Var[X+Y] = Var[X] + Var[Y] + 2COV[X,Y]$$



## Dealing with Dependent RVs: What is COV?

$$\begin{aligned} COV[X,Y] &= E[(X-\mu_x)(Y-\mu_y)] && (1)\\ &= E[XY] - E[X]E[Y] && (2)\end{aligned}$$

The connections between Eq 1 and Eq 2 is not obvious (and beyond this course).



## Standardizing COV

- It is convenient to standardize COV[X,Y] by adjusting for the size of the *spread* in each RV (X and Y).
- The standardized measure is called the *correlation coefficient*,

$$\rho_{x,y} = \frac{COV[X,Y]}{\sigma_x \sigma_y} \qquad (3)$$

$\rho$ (population parameter) is related to sample correlation $r$ (sample statistic) that will be discussed in a subsequent class.
Re-expressing Eq 3 we find:

$$COV[X,Y] = \rho_{x,y}\sigma_x\sigma_y$$



## Correlation Coefficient as Parameter

Note $\rho$ is bounded. $-1.0 \geq \rho \leq +1.0$.

$\rho_{xy} > 0$ i.e., positive implies that X and Y are positively correlated.

$\rho_{xy} < 0$ i.e., negative implies that X and Y are negatively correlated.



## Correlation and Z-Scores

We can re-express $\rho_{x,y}$ as a function of the z-scores for $X$ and $Y$.

$$\begin{aligned} \rho_{x,y} &= \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)P[x_i,y_i]}{\sigma_x\sigma_y} \\ &= \sum_{i=1}^{N}\frac{(x_i-\mu_x)}{\sigma_x}\frac{(y_i-\mu_y)}{\sigma_y}P[x_i,y_i] \\ &= \sum_{i=1}^{N} z_x z_y P[x_i,y_i] \end{aligned}$$

where $z_x$ denotes a z-score for $x$, and $P[x,y]$ denotes the probability that X takes on value x AND Y takes on value y.



## Summary: Single RV

- Adding a Constant ($c$)

$$\begin{aligned} \text{EV}\quad & E[X+c] = E[X] + c \\ \text{Var}\quad & Var[X+c] = Var[X] \\ \nu &= \frac{\sqrt{Var[X+c]}}{E[X+c]} = \frac{\sqrt{Var[X]}}{E[X]+c} \end{aligned}$$

- Multiplying by a constant ($a$)

$$\begin{aligned} \text{EV}\quad & E[aX] = aE[X] \\ \text{Var}\quad & Var[aX] = a^2 Var[X] \\ \nu &= \frac{\sqrt{Var[aX]}}{E[aX]} = \frac{a\sqrt{Var[X]}}{aE[X]} = \nu_x \end{aligned}$$

Where $\nu$ is defined as the coefficient of variation.

## Summary Two RVs (Independent)

EV $E[X+Y] = E[X] + E[Y]$ (applies always)

Var $Var[X+Y] = Var[X] + Var[Y]$ (iff independent)

$$\nu = \frac{\sqrt{Var[X+Y]}}{E[X+Y]} = \frac{\sqrt{Var[X]+Var[Y]}}{E[X]+E[Y]}$$

Where $\nu$ is defined as the coefficient of variation.



## Simple Case

Consider the simple case with no coefficients.

$$Z = X + Y$$



## Example

| | $\mu$ | $\sigma$ |
|---|---|---|
| X | 80 | 12 |
| Y | 20 | 3 |

- Consider $Z = X + Y$ when independent. Compute the $E[Z]$ and $Var[Z]$
- Now, how would answer change if $\rho_{xy} = 0.5$?
- Recall, only variance is affected by dependency.

$$\begin{aligned} Var[X+Y] &= \sigma_x^2 + \sigma_y^2 + 2\rho_{xy}\sigma_x\sigma_y \\ &= 12^2 + 3^2 + 2(0.5)(12)(3) \\ &= 144 + 9 + 24 \end{aligned}$$



## More General Case

$$Z = aX + bY$$

Note the coefficients $a$ and $b$. So for example we might have

- $Z = 0.2X + 0.15Y$
- $Z = 0.2X - 0.30Y$

In this case the general formula for variance is:

$$Var[Z] = a^2 Var[X] + b^2 Var[Y] + 2(ab)COV[X, Y]$$



## A Minus Sign is a Coefficient

Next, consider $Z = X - Y$

$$Var[Z] = \sigma_x^2 + \sigma_y^2 + (?)$$

Careful w/last part (?) ...
Recall that for $Z = aX + bY$, where the RVs $X$ and $Y$ are *dependent*.

$$\begin{aligned} Var[Z] &= a^2 Var[X] + b^2 Var[Y] + 2(ab)COV[X, Y] \\ &= a^2\sigma_x^2 + b^2\sigma_y^2 + 2(ab)\rho_{x,y}\sigma_x\sigma_y \end{aligned}$$



## A Minus Sign is a Coefficient

$Z = X - Y$
This can be re-expressed as,

$$(1)X + (-1)Y$$

, which indicates that $a = 1$ and $b = (-1)$.
*The key point is that* $b = -1$. Lets plug that into Eq 4 to get,

$$\begin{aligned} Var[Z] &= (1)^2\sigma_x^2 + (-1)^2\sigma_y^2 + 2(1)(-1)\rho_{xy}\sigma_x\sigma_y \\ &= 12^2 + 3^2 - 2(0.5)(12)(3) \\ &= \sigma_z^2 = 144 + 9 - 24 \end{aligned}$$

## More Examples: Independent

| | $\mu$ | $\sigma$ |
|---|---|---|
| X | 80 | 12 |
| Y | 20 | 3 |

First consider independent RVs
**Let** $Z = 0.25X + Y$

- $\sigma_z = \sqrt{0.25^2\sigma_x^2 + \sigma_y^2} = 4.24$

**Let** $Z = X - 0.5Y$

- $\sigma_z = \sqrt{\sigma_x^2 + 0.5^2\sigma_y^2} = \sqrt{144 + 9/4}$



## More Examples: Dependent

Now, how would answers change if $\rho_{xy} = 0.5$? Recall, only variance is affected by dependency.
**Let** $Z = 0.25X + Y$

$$\begin{aligned}\sigma_z &= \sqrt{0.25^2\sigma_x^2 + \sigma_y^2 + 2(0.25)\rho_{xy}\sigma_x\sigma_y} \\ &= \sqrt{144/16 + 9 + 2(0.25)(0.5)(12)(3)}\end{aligned}$$

**Let** $Z = X - 0.5Y$

$$\begin{aligned}\sigma_z &= \sqrt{\sigma_x^2 + 0.5^2\sigma_y^2 + 2(1)(-0.5)\rho_{xy}\sigma_x\sigma_y} \\ &= \sqrt{144 + 9/4 - 2(1)(0.5)(0.5)(12)(3)}\end{aligned}$$



## Special Case of Multivariate Normal

If the RV's $X$ and $Y$ follow and Normal probability model ($X, Y \sim$ N()) then any linear combination of these random variables (such as $X + Y$) will also follow a normal probability model.



## Example



## Area versus Inverse Problems

- Typical area problem follows the sequence:

$$X \Rightarrow Z \Rightarrow \text{Area} \Rightarrow \text{Probability}$$

- Typical inverse problem follows the reverse sequence:

$$\text{Probability} \Rightarrow \text{Area} \Rightarrow Z \Rightarrow X$$



## Inverse Problem: Example

$$\text{Probability} \Rightarrow \text{Area} \Rightarrow Z \Rightarrow X$$

Cereal Company problem from Sharpe (page 258-259).

## Types of Questions: A Summary

Four main types of questions:

Area Questions Probability that X lies in some range (e.g., $X > x_1$, $X < x_2$, or $x_1 < X < x_2$)

Expected value Uniform, Normal, Exponential

Variance calculation Uniform, Normal, Exponential

Inverse Problem Uniform, Normal (using tables), Exponential.

Continuous Probability Models & ...
└Summary

## Next Time

1. Back to data
2. Have already looked at categorical data ... now we examine quantitative data
   - Graphical displays for quantitative data
   - Describing the graphical displays
   - Quantitative summaries of quantitative data
3. Reading: Revisit Ch04 for review and read Ch05 ( Sharpe et al.)