Unit Schedule: Modules

Module	Week	Content	Ross
1.	1	Introduction to modelling for data science and to R	1,2
2.	2	Probabilities and bias	3
	3	Expectations	4
	4	**Distributions**	5
3.	5	Statistical inference	6&7
	6	Hypothesis testing	7&8
4.	7	Dependence and linear regression	9
	8	classification and clustering	
5.	9	Comparing means	10
	10	Random number generation and simulation	
6.	11	Validation and complexity	15
	12	Modelling	

FIT5197 Statistical Data Modelling

Module 2

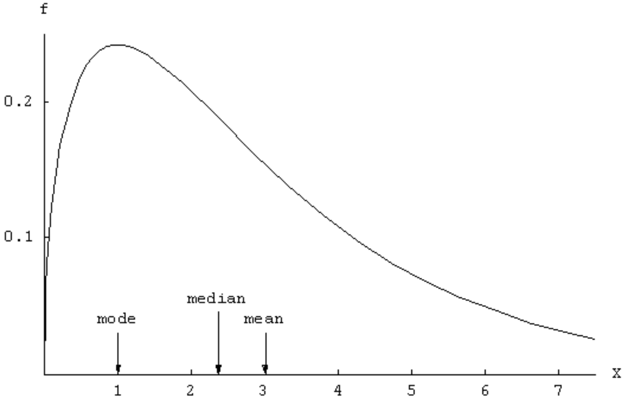
Special Distributions

2020 Lecture 4

Monash University

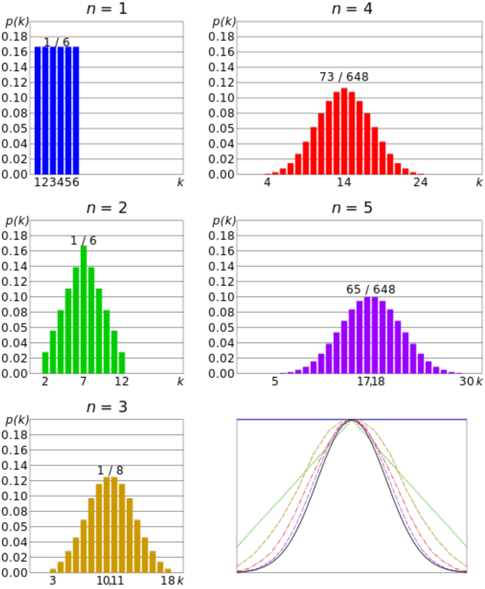
Revision at https://flux.qa/43FMK4

Refresher: Central Tendency



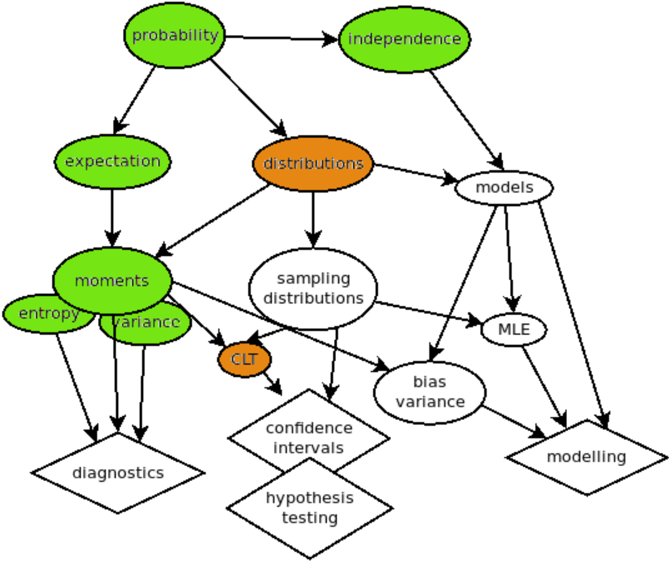
- in a skewed distribution, the mode, median and mean do not line up!
- long tail on right and hump on left means skewed to the right

Refresher: Summing Die



- sum of pips on n die
- $n = 1$ is uniform
- $n = 2$ is triangular
- n gets bigger, becomes more bell shaped
- distribution of the mean of large number of identical variables is usually Gaussian

Concept Map for This Unit



Special Distributions (ePub sections 2.3, Ross 5.1, 5.2, 5.4, 5.5)

Parametric Distributions

- so far we have built probability distributions from arguments or using particular shapes for $x \in \mathcal{X}$
- this is fine if $\mathcal{X}$ is a small finite set, or when you have particular reasons for the construction.
- there are several so-called parametric probability distributions that cover **useful/general classes of problems**
- we will look at several important distributions:
  - *Gaussian distribution*
  - *Bernoulli distribution*
  - *binomial distribution*
  - *Poisson distribution*
- in each case, we need to understand (1) the distribution and (2) the reasons/arguments for its use

Parametric Distributions, cont.

- we specify the probability density function by:

$$f(x|\vec{\theta}), \ \ x \in \mathcal{X}, \ \vec{\theta} \in \Theta$$

  or, for discrete RVs we use the shorthand notation:

$$P(X = x|\vec{\theta}) \equiv p(x|\vec{\theta}), \ \ x \in \mathcal{X}, \ \vec{\theta} \in \Theta$$

- where
  - $\vec{\theta} = (\theta_1, \ldots, \theta_k)$ are the parameters that control distribution of the probabilities
  - Θ is the set of valid parameters for the model
- the mean, the variance, cdf, quantiles, etc., are all computed from $f(x|\vec{\theta})$
  - these derived quantities appear in most text books and references

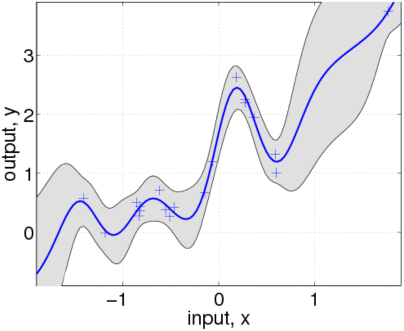
  e.g. see *beta distribution*

Outline

Gaussian Distribution

Motivation: Gaussian

Non-linear Multidimensional Curve Fitting



- example above is non-linear curve fitting with error bars
- done with the GPML Matlab code from *GaussianProcess.org*
- built using multivariate Gaussian theory
  - "Gaussian processes" – a non-parametric version

Gaussian Distribution, example

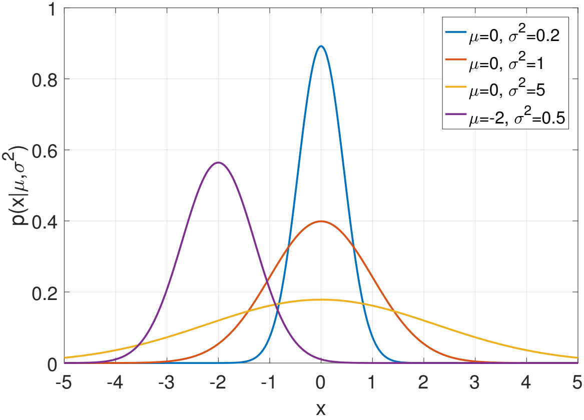


Figure: Probability density functions for several normal (Gaussian) distributions. Note that the normal distribution is symmetric and tails off to zero as $|x| \to \infty$.

The orange curve is the standard normal distribution.

Gaussian Distribution

- Let's begin with the case that $\mathcal{X} = \mathbb{R}$
  $\Longrightarrow$ that is, we want a distribution over all the real numbers
- Probably the most important distribution for real numbers is the Gaussian (normal) distribution
  $\Longrightarrow$ named after Carl Friedrich Gauss (1777-1855)
- the pdf for a Gaussian distribution is given by

$$f(x \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  where
  - μ is the mean of the distribution;
  - σ^2 is the variance of the distribution;

  so that $\vec{\theta} = (\mu, \sigma^2)$ for the Gaussian distribution.

Gaussian Distribution, cont.

- If X follows a Gaussian distribution, we write that

$$X \sim N(\mu, \sigma^2)$$

  where "$\sim$" is read as "is distributed per a"
- the case $N(0, 1)$ is called the standard normal distribution
- If $Z \sim N(0, 1)$, then

$$X = \sigma Z + \mu$$

  is distributed as per $N(\mu, \sigma^2)$

$\Longrightarrow$ Every Gaussian distribution is a translated and scaled version of the standard normal distribution $N(0, 1)$

Gaussian Properties

- If $X \sim N(\mu, \sigma^2)$, then

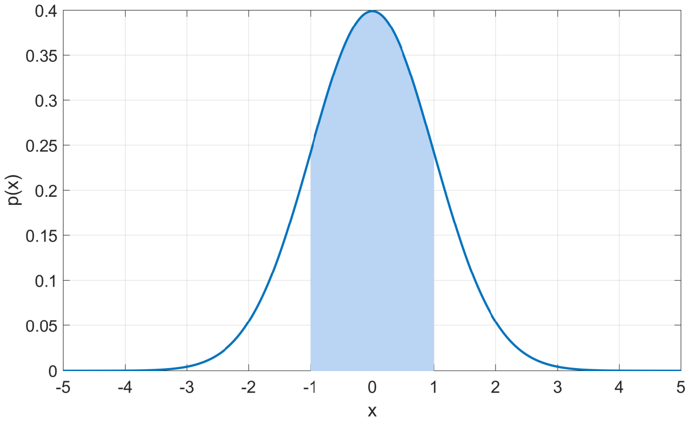
$$\begin{aligned} \mathbb{E}[X] &= \mu, \\ \mathbb{V}[X] &= \sigma^2. \end{aligned}$$

- the Gaussian distribution is symmetric around μ, so that:
  - its mode is μ;
  - its median is μ.
- the cdf for the Gaussian has no closed form
  - most packages have algorithms to evaluate it numerically
- many distributions converge to a Gaussian in some limit

  e.g. sample means when sample size $n \to \infty$

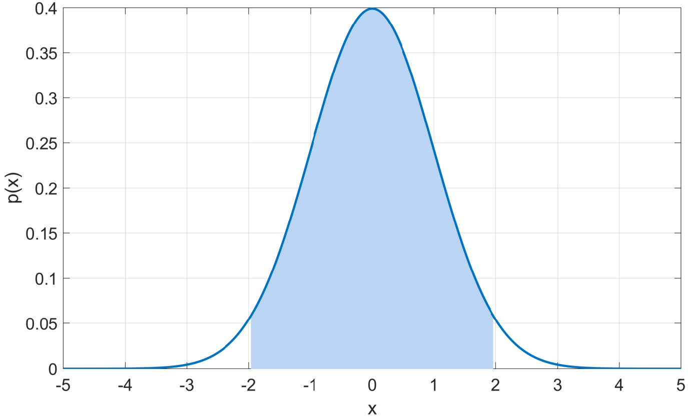
Gaussian Properties, cont.

- For any $N(\mu, \sigma^2)$:
  - 68.27% of probability falls within $(\mu - \sigma, \mu + \sigma)$



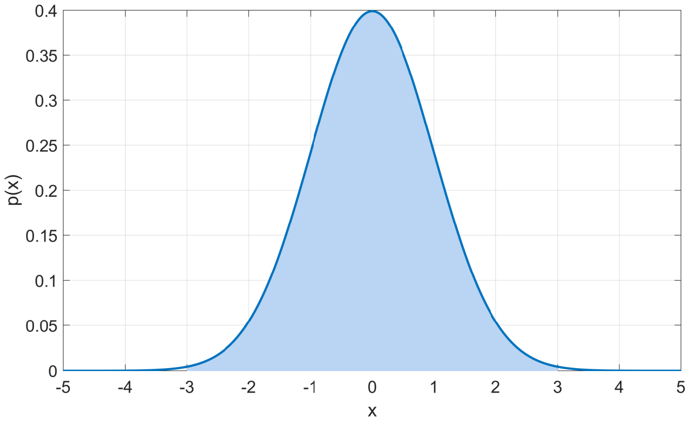
Gaussian Properties, cont.

- For any $N(\mu, \sigma^2)$:
  - 95.45% of probability falls within $(\mu - 2\sigma, \mu + 2\sigma)$



Gaussian Properties, cont.

- For any $N(\mu, \sigma^2)$:
  - 99.73% of probability falls within $(\mu - 3\sigma, \mu + 3\sigma)$

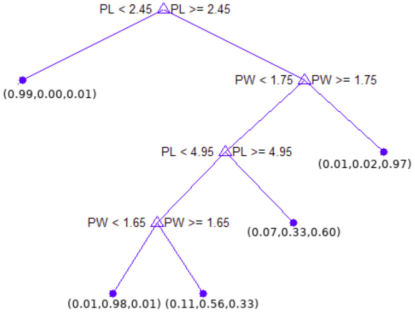


Outline

Bernoulli Distribution

Motivation: Bernoulli

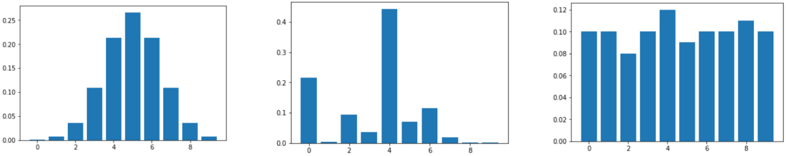
Rule-based Classification



- example above is a class probability tree built on the Iris data
- each node has a 3-class probability vector
- common with many classifiers

Motivation: Bernoulli

Multinomial Distribution



- these discrete distributions are extensions of Bernoulli's from 2 to $K = 10$ outcomes
- analysis is similar

Bernoulli Distribution

- the Bernoulli distribution models discrete, binary RVs, i.e., $\mathcal{X} = \{0, 1\}$

$$P(X = 1 \mid \theta) = \theta, \;\; \theta \in [0, 1]$$

so that the parametric probability distribution follows:

$$p(x \mid \theta) = \theta^x (1 - \theta)^{(1-x)}$$

e.g. tossing a single coin and looking for "heads"

- the parameter θ is the probability of observing a "success"
- If X follows a Bernoulli distribution, we write $X \sim \mathrm{Be}(\theta)$
- It is easy to see that

$$\begin{aligned} \mathbb{E}[X] &= \theta \\ \mathbb{V}[X] &= \theta(1 - \theta) \end{aligned}$$

Bernoulli Variance

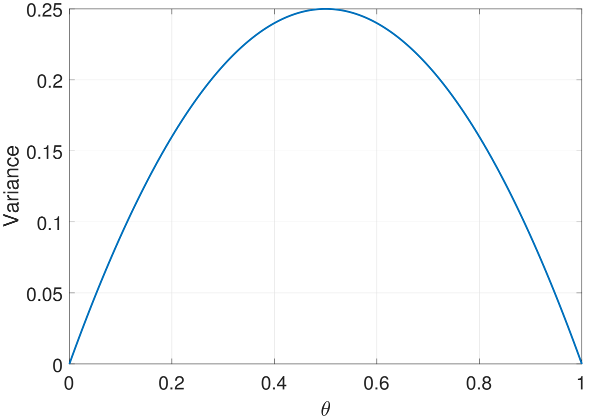


Figure: Variance of a Bernoulli random variable as a function of θ. The variance is maximum when $\theta = 1/2$ and smallest for $\theta = 0$ and $\theta = 1$.

Outline

Binomial Distribution

Motivation: Binomial

Bag of Words Model for Text

- generalisation of binomial distribution from 2 to K outcomes is called the multinomial distribution
  - but we study the binomial first
- convert a paragraph of text into a multinomial sample by ordering and counting

Original news article:	Despite their separation, Charles and Diana stayed close to their boys William and Harry. Here, they accompany the boys for 13-year-old William's first day school at Eton College on Sept. 6, 1995, with housemaster Dr. Andrew Gayley looking on.
Bag of words:	13 1995 accompany and(2) andrew at boys(2) charles close college day despite diana dr eton first for gayley harry here housemaster looking old on on school separation sept stayed the their(2) they to william(2) with year

Binomial Distribution

- Now consider n Bernoulli RVs $\mathbf{X} = (X_1, \ldots, X_n)$.
  - Example realisation: $\mathbf{x} = (0, 1, 1, 1, 0, 1, 0, 0, 1, 1)$
- the sum

$$m(\mathbf{x}) \equiv m = \sum_{j=1}^{n} x_j$$

  counts the number of "successes"
  $\Longrightarrow$ in our example, $m = 6$

e.g. tossing 10 coins and counting the total "heads"

N.B. binary version of tossing a dice 10 times and counting each face

- Given n, the count is a RV, say M, over the sample space $\{0, 1, 2, \ldots, n\}$

Binomial Distribution, example

- have $n = 4$ trials
- the following 6 sequences have $m = 2$ successes

$$\begin{array}{ccc} 1,1,0,0 & 1,0,1,0 & 1,0,0,1 \\ 0,1,1,0 & 0,1,0,1 & 0,0,1,1 \end{array}$$

- each of the 2 successes has probability θ
- each of the (4-2) failures has probability $(1 - \theta)$
- so that

$$P(m{=}2 \mid \theta) = 6\, \theta^2\, (1 - \theta)^{(4-2)}$$

How many ways of getting m successes in n trials are there?

Binomial Coefficient

The **binomial coefficient** denoted $\binom{n}{m}$ is given by

$$\binom{n}{m} = \frac{n!}{(n-m)!m!}$$

where $m! = 1 \times 2 \times 3 \times \ldots \times m$ is the factorial function.

- is the number of ways of choosing m objects out of n identical objects
- see *binomial coefficient*
- examples:

$$\binom{6}{1} = 6, \qquad \binom{6}{2} = 15$$

Binomial Coefficient

binomial coefficient is given by

$$\binom{n}{m} = \frac{n!}{(n-m)!m!} = \frac{n(n-1)\cdots(n-m+1)}{m!}$$

Explanation:

1. n choices for 1-st pick
2. $(n-1)$ choices for 2-nd pick

...

m. $(n-m+1)$ choices for m-th pick

then divide by $m!$ possible orderings of the m choices

Binomial Distribution, cont.

- the binomial distribution describes the probability that M takes a particular value m

$$p(m \mid n, \theta) = \binom{n}{m} \prod_{i=1}^{n} p(x_i \mid \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}$$

- this captures the fact that, for $1 \leq m \leq (n-1)$ there is multiple sequences with m successes out of n trials
- generalisation from 2 to K outcomes is called the *multinomial distribution* which is built on the *multinomial coefficient*

Binomial Distribution Plots

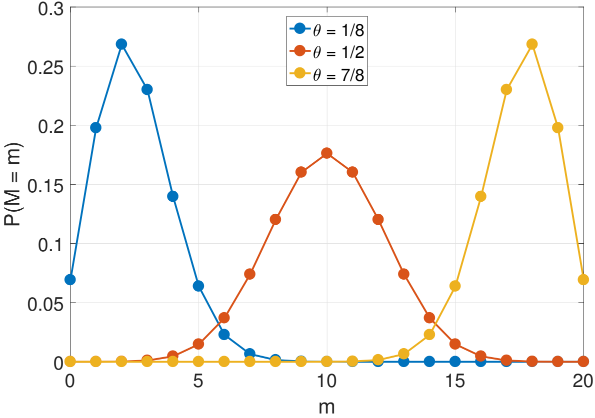


Figure: Binomial distribution for $n = 20$ and $\theta = 1/8$, $\theta = 1/12$, $\theta = 7/8$. The distribution is defined only on the integers – the connecting lines are only guides for the eye. Note that $\theta = 7/8$ is a mirror of $\theta = 1/8$.

Binomial Distribution, cont.

- If M follows a binomial distribution, we write

$$M \sim \mathrm{Bin}(\theta, n)$$

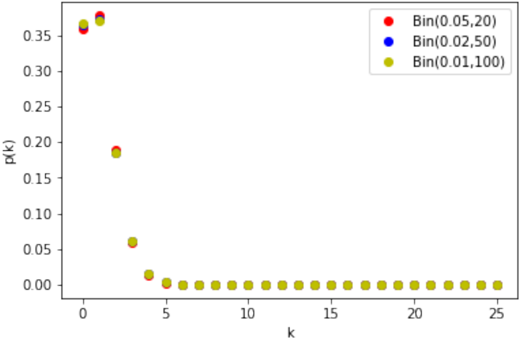
- As m is a sum of independent Bernoulli RVs we have

$$\begin{aligned} \mathbb{E}[M] &= n\theta \\ \mathbb{V}[M] &= n\theta(1-\theta) \end{aligned}$$

Using the Binomial

- the binomial distribution models the number of successful outcomes out of n total events
- when is the binomial appropriate?
  - the occurrence of one event does not affect the probability that a second event will be successful:
    - events occur independently
  - the probability at which events are successful is constant
  - a fixed number of events are recorded

Binomial, Large n, Small Mean



- curiously, the 3 binomial distributions are almost identical
- $n\theta$ = 20*0.05 = 50*0.02 = 100*0.01

when $n\theta$ is small and n large, the distribution can be approximated to one only dependent on $n\theta$

Binomial for Large n

Consider the probability for the binomial (proof optional)

$$
\begin{aligned}
p(x|n,\theta) &= \frac{n(n-1)\cdots(n-x+1)}{x!}\theta^x(1-\theta)^{n-x} \\
&= \frac{1}{x!}\left(\frac{\theta}{1-\theta}\right)^x n(n-1)\cdots(n-x+1)\,(1-\theta)^n \\
&\approx \frac{1}{x!}\left(\frac{n\theta}{1-\theta}\right)^x e^{n\log(1-\theta)} && \text{for } x \ll n \\
& && \text{(justified if } n\theta \text{ is small)} \\
&\approx \frac{1}{x!}(n\theta)^x\, e^{-n\theta} && \text{for } \theta \text{ small}
\end{aligned}
$$

$\Longrightarrow$ when n is large and $n\theta$ is small
distribution approaches $p(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$ for $\lambda = n\theta$

Outline

Poisson Distribution

Motivation: Poisson

Recommender System with Side Information

- user ratings ("stars") for books/videos can be viewed as Poisson data
- we want to estimate "propensity" for new users/books/videos based on side information

Poisson Distribution

- what if our data is non-negative integers; for example:
  - number of telephone calls made in an hour
  - number of people kicked to death by horses in a year
  - Sample space is then $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$
- one suitable distribution is the Poisson distribution
  - Named after Simeon Poisson (1781–1840)
- has the form

$$p(x \mid \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

  where λ is often called the *rate*

Poisson Distribution, Properties

- If X is distributed per a Poisson distribution we write

$$X \sim \mathrm{Pois}(\lambda)$$

- the Poisson distribution has

$$\begin{aligned} \mathbb{E}[X] &= \lambda \\ \mathbb{V}[X] &= \lambda \end{aligned}$$

- the Poisson distribution is an example of a distribution in which the variance grows with the mean
- corresponds to the large sample small mean case of the binomial

Poisson Distribution Plots

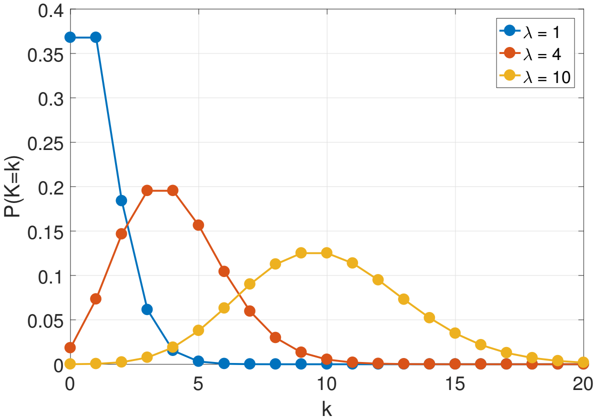


Figure: Poisson distribution for $\lambda = 1$, $\lambda = 4$ and $\lambda = 10$. The distribution is defined only on the integers – the connecting lines are only guides for the eye.

Using the Poisson

- the Poisson distribution models the number of events in an interval of time or space
- when is the Poisson appropriate? (taken from Wikipedia)
  - the occurrence of one event does not affect the probability that a second event will occur:
    - events occur independently
  - the rate at which events occur is constant:
    - the rate cannot be higher in some intervals and lower in other intervals.
  - two events cannot occur at exactly the same instant.
  - the probability of an event in a small interval is proportional to the length of the interval

Using the Poisson, example

- suppose n_1 is customers coming into shop 10am to 10:10am, $n_1 \sim \mathrm{Pois}(\lambda)$
- suppose n_2 is customers coming into shop 10:10am to 10:20am, $n_2 \sim \mathrm{Pois}(\lambda)$

$\Longrightarrow$ then by assumptions $n_1 + n_2 \sim \mathrm{Pois}(2\lambda)$

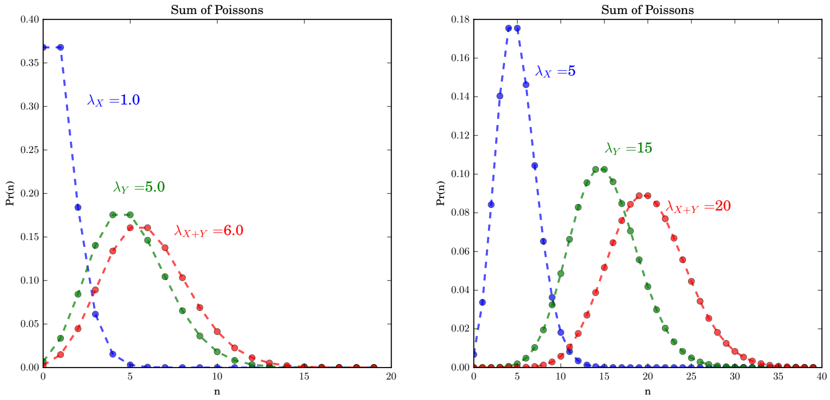
i.e. number of customers coming in between 10am and 10:20am **is distributed the same as** number of customers coming in between 10am and 10:10am at double the rate

- suppose n_3 is customers coming into shop 10:20am to 10:30am, $n_3 \sim \mathrm{Pois}(\lambda)$

$\Longrightarrow$ then by assumptions $n_1 + n_2 + n_3 \sim \mathrm{Pois}(3\lambda)$

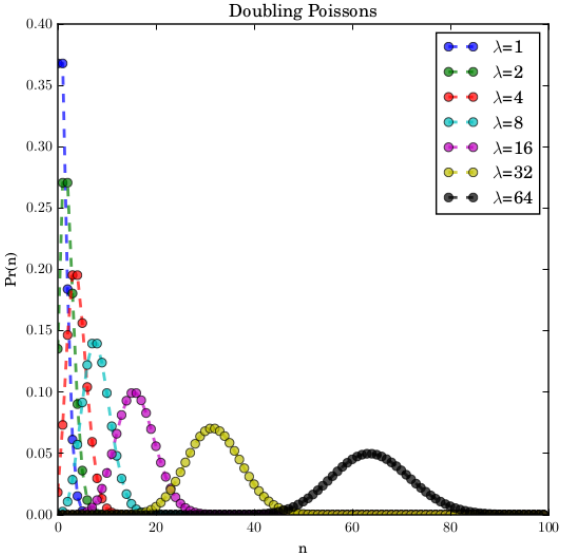
- Poisson distribution said to be divisible

Adding Poissons



if $X \sim \text{Pois}(\lambda_X)$ and $Y \sim \text{Pois}(\lambda_Y)$
then $(X + Y) \sim \text{Pois}(\lambda_X + \lambda_Y)$

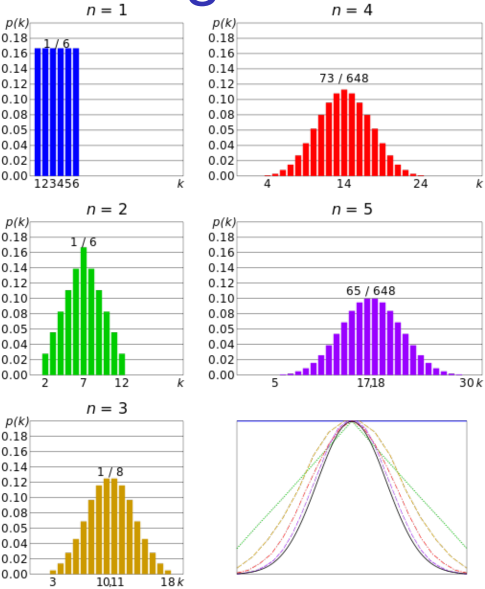
Doubling Poissons



Outline

Central Limit Theorem

Summing Die



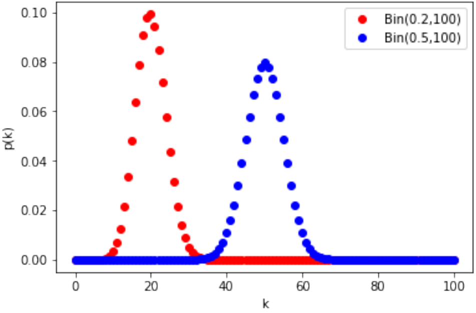
- sum of pips on n die
- let d_i be pips on i-th die, then

$$sum = \sum_{i=1}^{n} d_i$$

- n gets bigger, sum becomes more Gaussian
- note

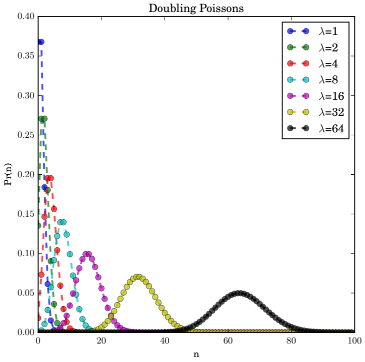
$$mean = \frac{1}{n} \sum_{i=1}^{n} d_i$$

Binomial Large n



- for $M \sim \mathrm{Bin}(\theta, n)$ have $M = \sum_{i=1}^{n} b_i$ for $b_i \sim \mathrm{Be}(\theta)$
- for large n, and mode not near the boundary, it becomes Gaussian
- for large n, and $n\theta$, $n(1-\theta)$ not too small, $M \sim \mathrm{Bin}(\theta, n)$ approaches $N(n\theta, n\theta(1-\theta))$

Dividing Poissons



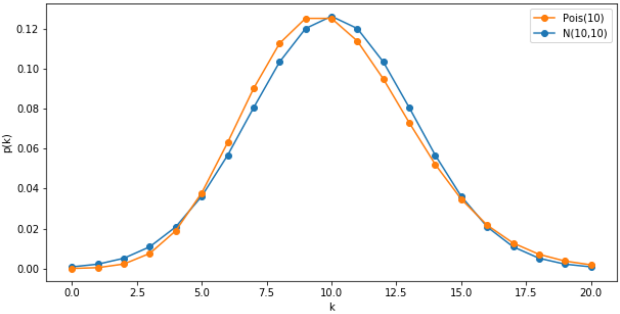
- consider $n \sim \mathrm{Pois}(\lambda)$ for $\lambda \gg 1$
- by divisibility it can be broken up into $n = n_1 + ..., n_K$ where each $n_k \sim \mathrm{Pois}(\lambda/K)$

e.g. $n \sim \mathrm{Pois}(64)$ is the sum of 64 variables of $\mathrm{Pois}(1)$

- for large $K < \lambda$, n can be made the sum of K identical variables, each not too skewed

$n \sim \mathrm{Pois}(\lambda)$ approaches $N(\lambda, \lambda)$ for large λ

Gaussian Approx. to Poisson



- $p(k|\mathrm{Pois}(10)) \approx p(k|N(10,10))$
- note Poisson is skewed to the right, whereas Gaussian is symmetric

Moments of the Mean

Have n identical RVs $X_1, ..., X_n$. Consider the mean $m_n = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\mathbb{E}[m_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X]$$

(proof optional)

$$\mathbb{V}[m_n] = \mathbb{E}\left[(m_n - \mathbb{E}[m_n])^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X])\right)^2\right]$$

$$= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}[X])^2 + \sum_{i \neq j = 1}^{n}(X_i - \mathbb{E}[X])(X_j - \mathbb{E}[X])\right]$$

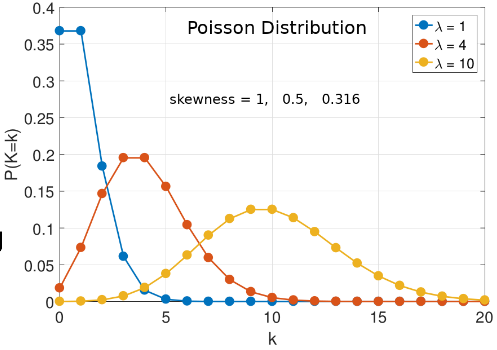
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mathbb{E}[X])^2\right] = \frac{1}{n}\mathbb{V}[X]$$

Skewness (optional)

Consider higher order moments. (Pearson's coefficient of) *skewness* is the 3rd central moment of the standardised variable:

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sqrt{\mathbb{V}[X]}}\right)^3\right]$$

NB. is positive when long tail on right



What happens to the skewness for the **sample mean**?

Skewness of the Mean (optional)

$$\mathbb{E}\left[(m_n - \mathbb{E}[m_n])^3\right] = \frac{1}{n^3}\mathbb{E}\left[\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X])\right)^3\right]$$

$$= \frac{1}{n^3}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}[X])^3 + 3\sum_{i\neq j=1}^{n}(X_i - \mathbb{E}[X])(X_j - \mathbb{E}[X])^2\right]$$

$$+ \frac{1}{n^3}\mathbb{E}\left[\sum_{i\neq j\neq k=1}^{n}(X_i - \mathbb{E}[X])(X_j - \mathbb{E}[X])(X_k - \mathbb{E}[X])\right]$$

$$= \frac{1}{n^3}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mathbb{E}[X])^3\right] = \frac{1}{n^2}\mathbb{E}\left[(X - \mathbb{E}[X])^3\right]$$

$\Longrightarrow$ so $\gamma_1(m_n) = \frac{1}{\sqrt{n}}\gamma_1(X)$

the skewness of the sample mean shrinks with factor with $\frac{1}{\sqrt{n}}$

Central Limit Theorem

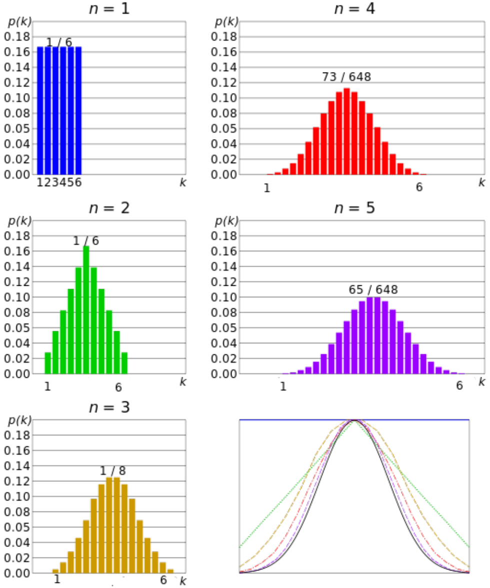
Central Limit Theorem (CLT):
Have distribution with mean μ and variance σ^2, and sample n identical RVs $X_1, ..., X_n$ from it. Then the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately distributed as $N\left(\mu, \frac{1}{n}\sigma^2\right)$ for large n.

Ross gives this differently:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \text{ is approximately standard normal}$$

- for distributions with higher skewness, convergence is slower!
- CLT fails to apply if mean or variance doesn't exist
- rough rule of thumb: best for $n > 30$

Central Limit Theorem: Die



- let d_i be pips on i-th die, then

$$mean = \frac{1}{n} \sum_{i=1}^{n} d_i$$

- n gets bigger, mean becomes more Gaussian

Outline

Distributions in R

Using Distributions in R

- every distribution has 4 functions:

  probability , `p`: is the cumulative density function (cdf)
  quantile , `q`: the inverse of cdf
  density , `d`: the the probability function; pdf or pmf
  random , `r`: Random number generator

- for Gaussian distributions: `pnorm()`, `qnorm()`, `dnorm()`, `rnorm()`

  e.g. `dnorm(x)` provides the Gaussian density
  e.g. `pnorm(x)` provides $\int_{-\infty}^{x} dnorm(x) \mathrm{d}x$
  e.g. `qnorm(p)` provides p-th quantile, x such that `pnorm(x)=p`, so `qnorm(0.25)` is the first quartile
  e.g. `rnorm(x)` does a random sample according to `pnorm(x)`

Distributions in R

Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank

End of Week 4