Methods of Mathematical Statistics

Notes by Tim Brown and Guoqi Qian

Module 2: Discrete Distributions

Contents

1 Discrete Random Variables — 2.1

1.1 Where from, Where to and Definitions

Where from?

Rules of probability and assumptions about probabilities or conditional probabilities dictate solutions for simple probability problems.

Assumption of equally likely outcomes gives probabilities and conditional probabilities but requires clear definition of the sample space.

Conditional probabilities or independence can instead be a key assumption.

Bayes' theorem needed to reverse conditional probabilities — used in statistics.

Where to?

Okam's razor says we should seek the simplest explanation that fits the facts.

Often the results of our experiment or observation is a number or numbers and the events of interest centre on these number(s).

For example, in the rocket failure example, we might only be interested in the number of components that fail.

And in the soft drink example, we might only be concerned about the number of chains that give more shelf space to P.

And in the sampling lines of code example, we might only be concerned about the total number of lines of code that could be improved.

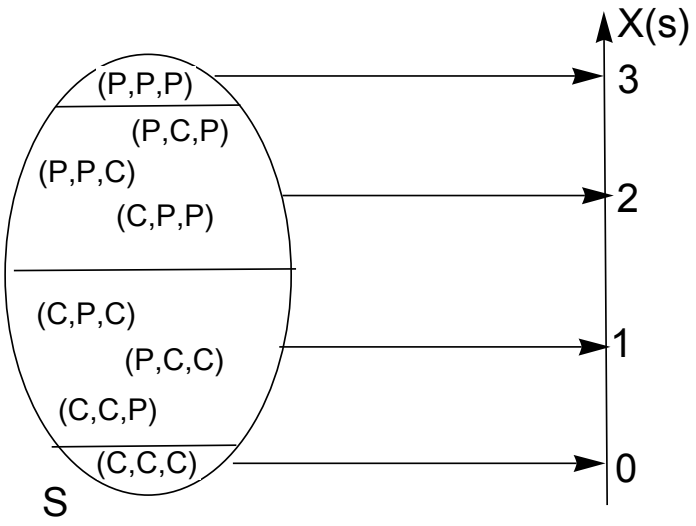

Figure 1: RV X, the number of chains with more shelf space for P

Definitions

Random Variable: Given an outcome space S, a function or rule, X, which associates with each $s \in S$ a number $X(s)$ is called a *random variable*, abbreviated *RV* or *rv*.

Range of a RV: The *range* or (Hogg and Tanis) *space* of a rv X is the set of real numbers achieved by X, i.e., $\{x : X(s) = x \text{ for some } s \in S\}$.

Abbreviation If B is a set of numbers, the notation $[X \in B]$ or $\{X \in B\}$ is an abbreviation for $\{s : X(s) \in B\}$ and $P(X \in B)$ for $P(\{s : X(s) \in B\})$.

The sample space may take a back seat in describing RV's if what matters are the probabilities associated with events defined by X.

A *discrete RV* is one with a finite range or an infinite range like $1, 2, \ldots$.

1.2 Example: Soft Drinks

Illustration: Soft Drinks

Illustration: Soft Drinks

The probabilities for X are:

3 chains

$$P(X = 3) = P(\{(P, P, P)\}) = \frac{1}{8}$$

2 chains

$$P(X = 2) = P(\{(P, P, C), (P, C, P), (C, P, P)\}) = \frac{3}{8}$$

1 chain

$$P(X = 1) = P(\{(P, C, C), (C, C, P), (C, P, C)\}) = \frac{3}{8}$$

0 chain

$$P(X = 0) = P(\{(C, C, C)\}) = \frac{1}{8}$$

Example: Soft Drinks rv

Continuing: If X is the number of supermarket chains that give more shelf space to soft drink P, what is the probability that X is even? odd?

Even:

$$P(X \text{ is even}) = P(X = 0) + P(X = 2) = \frac{4}{8}$$

Odd:

$$P(X \text{ is odd}) = P(X = 1) + P(X = 3) = \frac{4}{8}$$

1.3 Probability Mass Function

Probability Mass Function for Soft Drinks X

All the probabilities to do with X are given by $P(X = 0), P(X = 1), P(X = 2), P(X = 3)$ because $P(X = x) = 0$ for any number x outside the range $\{0, 1, 2, 3\}$.

For this reason, the rule which gives for a number $x = 0, 1, 2, 3$ the probability $P(X = x)$ is called the *the probability mass function* for the random variable.

Small letters are used for values in the range (i.e., numbers) and capital letters for the names of random variables.

General Definition

As long as the discrete rv X is the only quantity of interest in our experiment or observation, the probability mass function is sufficient for calculations.

The *probability mass function* abbreviated *pmf* or *PMF*, $f(x)$, of real numbers, x, is a function satisfying:

(a) $f(x) > 0$ for $x \in \text{range}(X)$ and $f(x) = 0$ for $x \notin \text{range}(X)$,

(b)

$$\sum_{x \in \text{range}(X)} f(x) = 1,$$

for any set B of numbers,

(c)

$$P(X \in B) = \sum_{x \in \text{range}(X) \cap B} f(x).$$

Rocket Failure: PMF Example

What is the probablity mass function (pmf) for the number of rocket components that do not fail?

Rocket Failure PMF: Solution

Let X be the random variable giving the number of rocket components that do not fail.

The range of X is $\{0, 1, 2, 3\}$ and recall that A_1, A_2, A_3 are the events that components 1, 2 and 3 fail.

By independence

$$\begin{aligned} P(X = 0) &= P(A_1 \cap A_2 \cap A_3) \\ &= P(A_1)P(A_2)P(A_3) = 0.15^3 = 0.00338. \end{aligned}$$

Splitting $[X = 1]$ into the three mutually exclusive events that partition it. Using the addition rule and independence gives

$$\begin{aligned} P(X = 1) &= P(A_1^c \cap A_2 \cap A_3) + P(A_1 \cap A_2^c \cap A_3) + P(A_1 \cap A_2 \cap A_3^c) \\ &= 3 \times 0.15^2 \times 0.85 = 0.057375. \end{aligned}$$

Rocket Failure PMF: Solution Ctd

Similarly

$$\begin{aligned} P(X = 2) &= P(A_1^c \cap A_2^c \cap A_3) + P(A_1 \cap A_2^c \cap A_3^c) + P(A_1^c \cap A_2 \cap A_3^c) \\ &= 3 \times 0.15 \times 0.85^2 = 0.325125 \\ P(X = 3) &= P(A_1^c \cap A_2^c \cap A_3^c) \\ &= 0.85^3 = 0.614125 \end{aligned}$$

The probability mass function can be written as a table:

x	0	1	2	3
P(X=x)	0.003375	0.057375	0.325125	0.614125

Rocket Failure PMF: Solution Ctd 2

Alternatively, the probability mass function can be viewed as a dot plot:

Rocket Failure PMF: Solution Ctd 3

Or the probability mass function can be viewed as a probability histogram:

Rocket Failure PMF: Comment

The probability computed in the first Rocket Failure example in Module 1 was $P(X \geq 1)$, which was computed as $1 - P(X = 0)$.

An alternative would have been to compute it as $P(X = 1)+P(X = 2)+P(X = 3)$ but this is much more complicated.

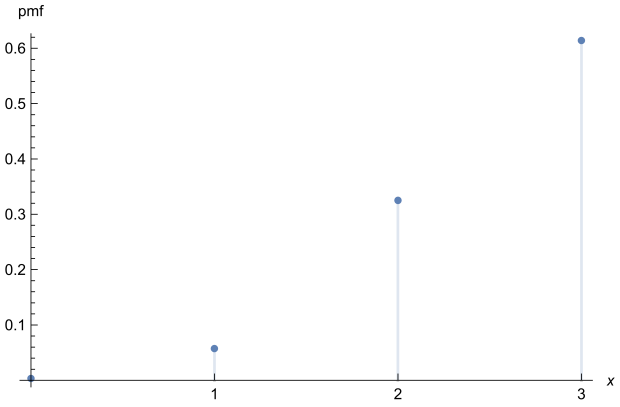

Figure 2: Probability Mass Function for X

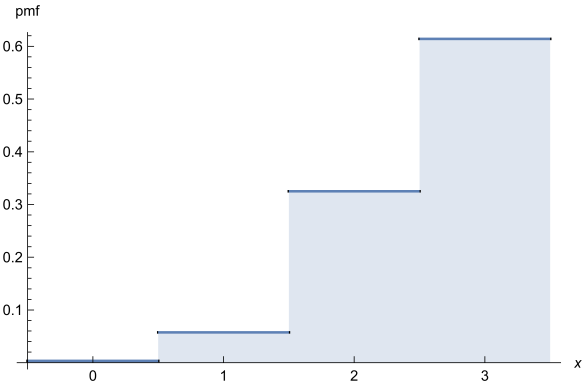

Figure 3: Probability Histogram for pmf: shaded blue area = 1

1.4 Binomial Distribution (Start of Ch 2.4)

Similarities

Similar Examples: In both the soft drink example and the rocket failure example, the pmf calculations for the rv X were similar.

Difference: The only difference was the probability 0.5 in the soft drink example (equally likely outcomes) compared to 0.85 in the rocket failure example.

Structure: Both examples had the idea of a *successful* outcome in each of 3 **independent** *trials*.

Soft Drink Example: The *trials* were the decisions of the three supermarket chain and *success* was soft drink P having more shelf space.

Rocket Example: The *trials* were the performance of the components and *success* was a component not failing.

RV of interest: In both cases, the rv, X, was the *number* of successes.

Bernoulli Trials Definition

Bernoulli Trials: Consider a number, n, of **independent** *trials*.

Two Outcomes: Each trial has one of two possible outcomes: *success* or *failure* denoted S and F.

Events of success: The results of the trials define **independent** events $A_1, A_2, \ldots, A_n$ representing success, S, on trials $1, 2, \ldots, n$.

Events of failure: So the events of failure, F, on the same trials are $A_1^c, A_2^c, \ldots, A_n^c$.

Binomial RV X

Binomial RV: Consider the rv, X, which is the number of trials that result in success.

Range: The range of X is $\{0, 1, 2, \ldots, n\}$.

Component RVs: For $i = 1, 2, \ldots, n$, let X_i be the rv which is 1 on the event A_i and 0 on A_i^c, so that X_i is 1 for a success on trial i and 0 for a failure on trial i.

So: The rv X_i *counts* the *number* of successes on trial i.

Now: The *number* of successes overall is X and this is the sum of the number of successes on each trial so

$$X = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i. \tag{1}$$

1.5 Derivation of Binomial Probabilities

PMF of Binomial RV — Informal

Probability of Success: Let p be the constant probability of success on each trial.

Component probs: By independence, each way to get x successes in the n trials has probability $p^x(1-p)^{n-x}$.

Patterns of S and F: The different ways to get x successes are just selections of the trial numbers on which there are to be successes — each of the remaining trials must have a failure.

Number of patterns: There are $\binom{n}{x}$ ways to get x successes in n trials.

Hence For $x = 0, 1, 2, \ldots, n$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{2}$$

PMF of Binomial RV — More Formal

S trials and F trials: Suppose $\{i_1, i_2, \ldots, i_x\}$ is a selection of x trials from $\{1, 2, \ldots n\}$ to be S and that $\{j_1, \ldots, j_{n-x}\} = \{i_1, i_2, \ldots, i_x\}^c$ are the trials to be F

Independence gives

$$P(A_{i_1} \cap \cdots \cap A_{i_x} \cap A^c_{j_1} \cdots \cap A^c_{j_{n-x}}) = p^x (1-p)^{n-x}.$$

Number of subsets: There are $\binom{n}{x}$ subsets of size x from $\{1, 2, \ldots, n\}$, and each subset is a choice of $\{i_1, i_2, \ldots, i_x\}$.

The corresponding events of the pattern of success and failure, $A_{i_1} \cap \cdots \cap A_{i_x} \cap A^c_{j_1} \cdots \cap A^c_{j_{n-x}}$ are disjoint.

Hence For $x = 0, 1, 2, \ldots, n$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

1.6 Sampling with Replacement

Example: Sampling with Replacement

There is a probability of 0.9 that each member of the class has worked outside class on this subject on the day before any particular lecture day. A spinner chooses one of the numbers from 1 up to 21 randomly to find out whether that member of the classes worked on this subject yesterday. This is repeated on 4 lecture days in total. What is the probability mass function for the random variable X which gives the number of class members chosen who worked on this subject on the day prior to the one when they were chosen?

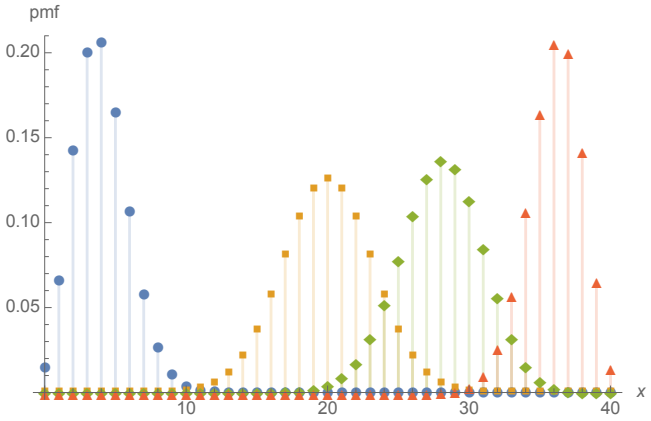

Figure 4: PMF's for n = 40, p=0.1,0.5,0.7,0.9

Solution: Sampling with Replacement

There are 4 Bernoulli trials with success being that the class member chosen worked on this subject on the day prior to them being chosen.

The trials are independent since the spinner choices are independent and we are told the probability is constant.

Hence,

$$P(X=0) = 0.1^4 = 0.0001$$
$$P(X=1) = 4 \times 0.9 \times 0.1^3 = 0.0036$$
$$P(X=2) = \frac{4 \times 3}{2} \times 0.9^2 \times 0.1^2 = 0.0486$$
$$P(X=3) = 4 \times 0.9^3 \times 0.1 = 0.2916$$
$$P(X=4) = 0.9^4 = 0.6561$$

1.7 Pictures of Binomial Probabilities

Binomial Probablities

1.8 Hypergeometric — Sampling without Replacement

Hypergeometric — Sampling without Replacement

Example: In the example of improving lines of code in the computer, a sample *without* replacement was taken at random.

The RV of interest was the total number, X, of lines of code in the sample that could be improved.

Notation: Given a *total* number, t, of lines of code.

Notation: a sample size, n, for the sampled lines of code.

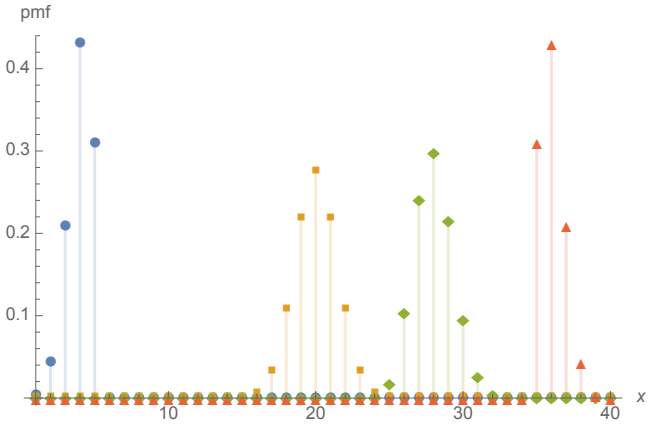

Figure 5: PMF's n = 40, p=0.1,0.5,0.7,0.9, t=50

Notation: a *total* number, b, of lines of code that can be improved.

Hypergeometric PMF: The same reasoning as the example gives, for $x = 0, 1, 2, \ldots, n$

$$P(X = x) = \frac{\binom{b}{x}\binom{t-b}{n-x}}{\binom{t}{n}}. \tag{3}$$

Note the top (and bottom) lines of the combination numbers in the numerator add to the top (resp. bottom) line in the denominator.

1.9 Hypergeometric — Derivation

Hypergeometric — Derivation

Samples: Each sample is a subset of size n from the t lines of code, so there are $\binom{t}{n}$ samples.

Sample lines improved:There are $\binom{b}{x}$ ways to choose the x lines in the sample from the b lines that can be improved.

Sample lines not improved: There are $\binom{t-b}{n-x}$ ways to choose the $n - x$ lines in the sample from the $t - b$ lines that can't be improved.

Multiplication principle says that the total number of ways to choose both the lines that can, as well as the lines that can't, be improved is the multiplication of the two numbers.

Equally likely random samples now completes the derivation

1.10 Hypergeometric — Pictures

Hypergeometric Pictures

Hypergeometric Pictures — Difference?

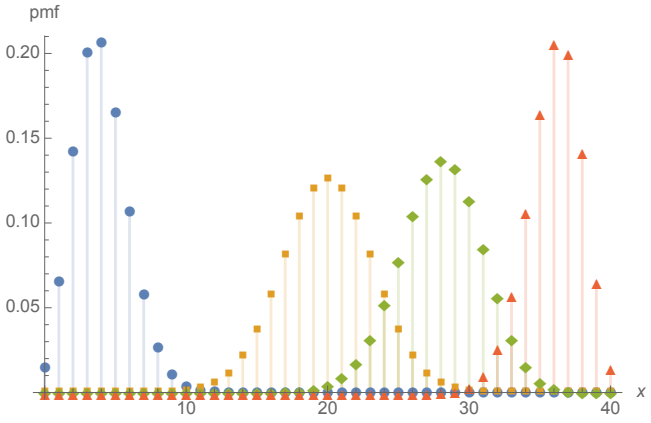

Figure 6: Same n and p but t = 10,000

Example: Capture-recapture experiment

Ten fish have been captured, tagged, and released to mix into their population. Suppose the population consists of 80 fish. A new sample of 15 animals is selected at random. What is the probability that, in the new sample, 3 will come from the tagged group?

Capture-recapture

Let X be the number of tagged animals in the new sample.

Then X has the hypergeometric distribution with $t = 80$, $n = 15$ and $b = 10$.

Therefore

$$P(X = 3) = \frac{\binom{10}{3}\binom{70}{12}}{\binom{80}{15}} = 0.1924.$$

Calculations like this are used in practice. Unlike the code example, it is the total population size, t, about which inference is needed. Bayes Theorem can be used. Governments use these, or alternative frequentist, techniques for monitoring population sizes and making policy decisions.

2 Expectation and Variance — 2.2, 2.3

2.1 Mean and variance of data — Section 6.1

Example: Travel Time

My travel time to work varies according to whether I drive or take public transport and also on the time of day. A typical travel time is 45 minutes by tram and bus but it can be as little as 20 minutes by car early in the morning. Over a two week period, my travel times were 45 min. ± a *departure time (min.)* given by the following table:

Table of Departures from 45 Minutes

Day	1	2	3	4	5	6	7	8	9	10
Dep. (min.)	+5	-5	-5	+5	+5	-25	+10	-25	-5	-20

Find the mean and empirical variance of the departures from my standard travel time of 45 minutes.

Definitions

For data $x_1, x_2, \ldots, x_n$ the *mean* is $\bar{x}$ given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{4}$$

and the *empirical variance* is v given by

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}. \tag{5}$$

Solution: Travel Time

For our departures from the standard travel time of 40 minutes we have the mean or average departure as

$$\bar{x} = \frac{5 - 5 + \cdots - 20}{10} = -6. \tag{6}$$

and the *Empirical Variance* is v given by

$$v = \frac{(5+6)^2 + (-5+6)^2 + \cdots + (-20+6)^2}{10} = 15.4. \tag{7}$$

2.2 Definition of Expectation of RV

Definition of Expectation

Random Selection: Consider the experiment of randomly selecting one departure time out of the 10 with equal chance for each of the 10 days.

Sample Space: The sample space is $1, 2, ..., 10$ and let X be the random variable which records the departure from 45 minutes for the randomly selected day.

Mean in Ex. : The number $\bar{x} = -6$ is then called the *Expectation* or *Mean* of the random variable X and denoted $E(X)$.

General: For any random variable X which is discrete

$$E(X) = \sum_{s \in S} X(s) P(\{s\}). \tag{8}$$

Definition of Expectation

Repetitions: In computing the sum, we could collect together all the outcomes s which have the same value of $X(s)$.

Travel Times: there were three repetitions of -5, three reps. of +5 and 2 reps. of -25.

General: Collecting outcomes together with the same value of the random variable gives:

$$E(X) = \sum_{x \in range(X)} xP(X = x). \tag{9}$$

2.3 Game Example

Example: Game

In a simple game, a die is rolled and 1¢ is paid if the roll of the die gives 1,2 or 3, 5¢ for a roll of 4 or 5 and 35¢ for a 6. What is a fair charge for this game?

Solution: Game

Definition: Let Y be the random variable which gives the prize on one play of the game.

Long Run Frequencies: Over many plays of the game, the prize will be y¢ (y = 1, 5 or 35) with approximate fraction $P(Y = y)$ of the time.

Long Run Gain: Hence, adding over a long sequence of games, the average amount of money (in ¢) I win will get closer to

$$1 \times P(Y = 1) + 5 \times P(Y = 5) + 35 \times P(Y = 35) = E(Y).$$

Logical Fair Charge is this average amount of money.

Solution: Game Ctd

PMF:

$$P(Y = 1) = \frac{3}{6}, P(Y = 5) = \frac{2}{6}, P(Y = 35) = \frac{1}{6}.$$

Answer:

$$E(Y) = 1 \times \frac{3}{6} + 5 \times \frac{2}{6} + 35 \times \frac{1}{6} = 8$$

i.e., a fair charge is 8c.

Comments: Game

Actual electronic games machines operate like this with more complicated chance mechanisms and more elaborate prize scales.

Legislation often limits the charge to the expected return (that is the expectation just calculated) plus a defined profit margin.

Example: In New South Wales clubs, the return is legislated to be at least 85% of the money input.

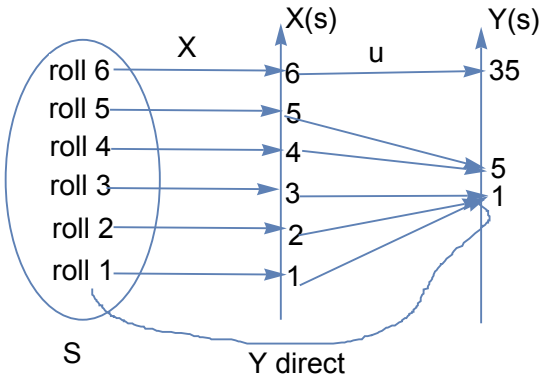

Figure 7: Y is a composition of u with X

2.4 Expectation of a Transform

Game: Another Solution

Start with a different random variable, X, which is the number on the die so that X has probability mass function $P(X = x) = \frac{1}{6}$ for $x = 1, 2, \ldots, 6$.

RV Y: is obtained by applying the function u to X so that

$$Y = u(X)$$

where

$$u(1) = u(2) = u(3) = 1,\ u(4) = u(5) = 5,\ u(6) = 35.$$

Diagram of Y= u(X)
Figure 7 shows X, Y.

Expectation of Y
This can be equally found as:

$$E(Y) = \sum_{s \in S} Y(s) P(\{s\}) \tag{10}$$

$$= \sum_{x=1}^{6} u(x) P(X = x) \tag{11}$$

$$= \sum_{y=1,\ 5 \text{ or } 35} y P(Y = y) \tag{12}$$

Different Levels of Agreggation

Aggregation: at three levels in the three different ways in equations (10), (11), (12) to compute $E(Y)$.

1 to 1: In this example, there was no aggregation from the sample space S to the range of X — one to one correspondence between outcomes in the sample space and results for Y.

But would not have had this if the sample space had two throws of the dice and X recorded the outcome of the first throw — the 36 sample points aggregated to 6 outcomes for X.

Aggregation: from 6 outcomes in range of X via the function u to 3 prize values for Y.

Different Levels of Agreggation

RHS's the same in equations (10), (11), (12) by the distributive law.

Textbook uses the second one, (11), as the definition.

PMF of X known means the second one, (11), is a good choice.

Need

$$E(|u(X)|) = \sum_x |u(x)|P(X = x) < \infty$$

to be safe when there are infinitely many values for a random variable (see the section in this chapter on Negative Binomial for an example).

Another Example on Aggregation/Composition

Suppose the sample space has 4 outcomes, $S = \{s_1, s_2, s_3, s_4\}$.

Suppose further that $P(\{s_i\}) = i/10,\ i = 1, 2, 3, 4.$

Note that these probabilities add to 1 because $1 + 2 + 3 + 4 = 10$.

Suppose finally that the random variable X is defined by $X(s_1) = X(s_2) = 1, \quad X(s_3) = -1, \quad X(s_4) = -2.$

Illustrate X and the random variable $Y = X^2$ on a diagram as a composition of X and the function $u(x) = x^2$.

Find $E(Y)$ in three different ways and order the ways in terms of aggregation.

Pictures for Aggregation/Composition, X

Figures 8, 9, 10 and 11 illustrate X, Y.

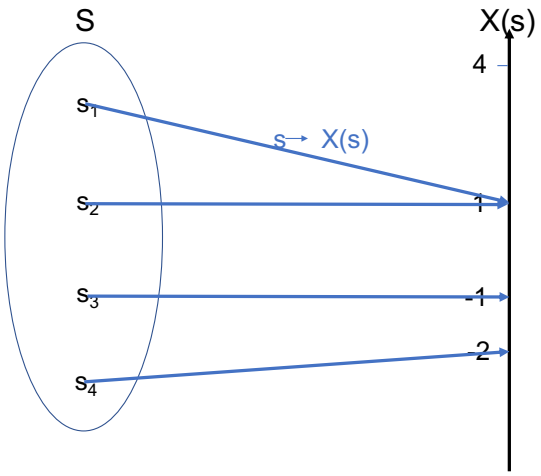

Figure 8: The random variable X in blue

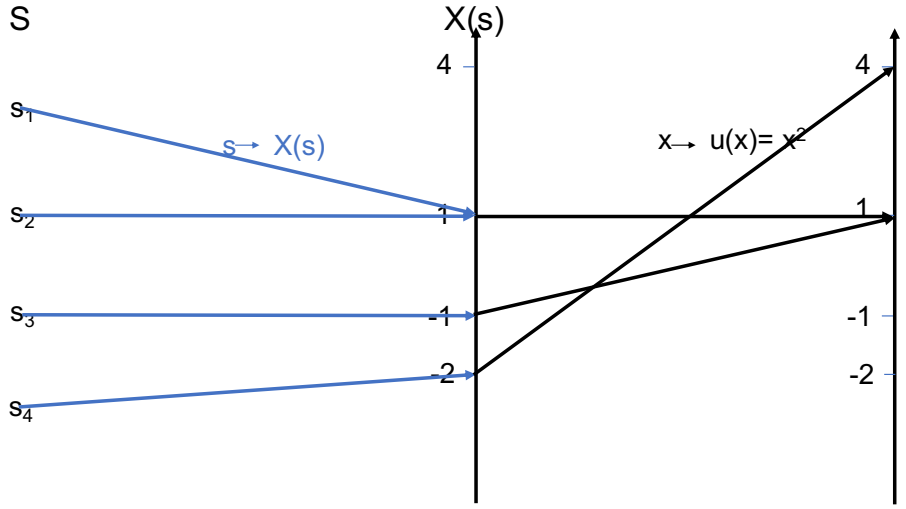

Figure 9: Adding the function u in black

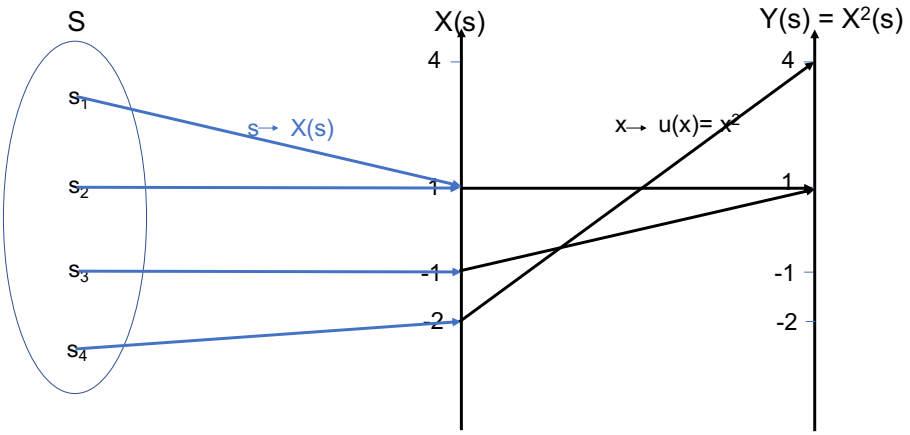

Figure 10: Y as a composition of $u(x) = x^2$ with X

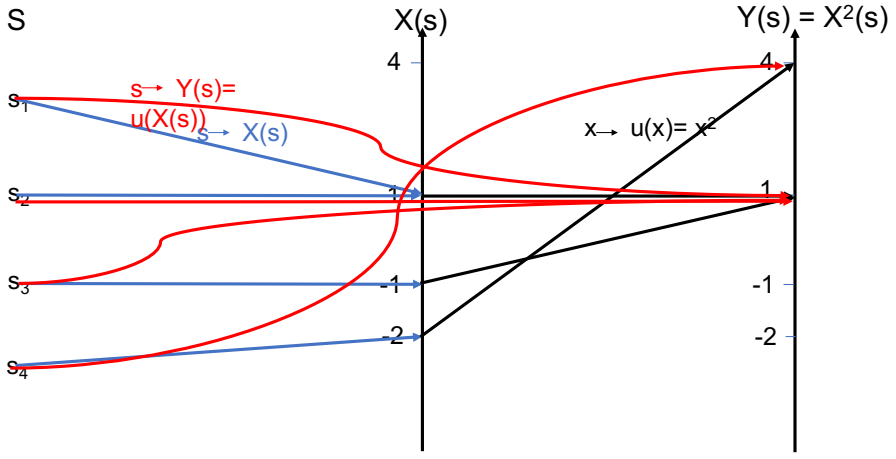

Figure 11: Y directly defined in red

Another Example: 3 different calculations

1. Sample space calculation:

$$\begin{aligned} E(Y) &= Y(s_1)P(\{s_1\}) + Y(s_2)P(\{s_2\}) + Y(s_3)P(\{s_3\}) + Y(s_4)P(\{s_4\}) \\ &= 1 \times \frac{1}{10} + 1 \times \frac{2}{10} + 1 \times \frac{3}{10} + 4 \times \frac{4}{10} = 2.2 \end{aligned}$$

2. Calculation using pmf of X and $u(x) = x^2$:

$$\begin{aligned} E(Y) &= \sum_{x \in range(X)} u(x)P(X = x) \\ &= u(1)P(\{s_1, s_2\}) + u(-1)P(\{s_3\}) + u(-2)P(\{s_4\}) \\ &= 1^2 \times (\frac{1}{10} + \frac{2}{10}) + (-1)^2 \times \frac{3}{10} + (-2)^2 \times \frac{4}{10} = 2.2 \end{aligned}$$

Another Example: 3 different calculations

3. Calculation using pmf of Y:

$$\begin{aligned} E(Y) &= \sum_{y \in range(Y)} yP(Y = y) \\ &= 1 \times P(\{s_1, s_2, s_3\}) + 4 \times P(\{s_4\}) \\ &= 1 \times (\frac{1}{10} + \frac{2}{10} + \frac{3}{10}) + 4 \times \frac{4}{10} = 2.2 \end{aligned}$$

There is more aggregation in Calculation 2 than in Calculation 1, and more aggregation in Calculation 3 than in Calculation 2.

2.5 Variance of a RV

Definition of Variance of a RV

Travel Time Ex. had *Empirical Variance* defined as the average of squared deviations around the mean:

$$v = \sum_{x=1}^{10} (x_i - \bar{x})^2 \times \frac{1}{n}\,.$$

Mean of RV: $E(X) = \mu$ (say), then μ is just some number depending on PMF of X — see (9).

Squared deviations from μ are measured by the function $u(x) = (x - \mu)^2$.

RV analogue of *Empirical Variance* is the *Variance*, often written σ^2, of the RV defined by

$$\begin{aligned} Var(X) &= E(u(X)) && (13) \\ &= E\big[(X - \mu)^2\big] = E\big[\big(X - E(X)\big)^2\big] && (14) \\ &= \sum_{x \in range(X)} (x - \mu)^2 P(X = x). && (15) \end{aligned}$$

Example: Variance of Game

Find the variance of the random variable Y in the simple game with the die.

Solution: Variance of Game

From the previous example, we found $E(Y) = 8$.

So

$$Var(Y) = (1-8)^2 \times \frac{3}{6} + (5-8)^2 \times \frac{2}{6} + (35-8)^2 \times \frac{1}{6} = 149.$$

Expectation and Variance in one pass

$\mu = E(X)$ must be calculated first to *define* the function $u(x) = (x-\mu)^2$ and hence to calculate the variance of the random variable.

Properties of expectation will enable the calculation of variance at the same time as the calculation of the mean.

The needed properties are covered in the next slides.

2.6 Expectation Linear, i.e., Distributive

Properties of Expectation

Constant: If c is a constant, then $E(c) = c$.

Multiplication by a constant: If c is a constant and X is a random variable, then

$$E(cX) = cE(X). \tag{16}$$

Additivity: If X and Y are random variables then

$$E(X+Y) = E(X) + E(Y). \tag{17}$$

Linearity: If $c_1, c_2, \ldots, c_n$ are constants and $X_1, X_2, \ldots, X_n$ are random variables, then

$$E\big(\sum_{i=1}^{n} c_i X_i\big) = \sum_{i=1}^{n} c_i E(X_i). \tag{18}$$

Demonstration

Constant: Using the definition of expectation in (8), the fact that probabilities of sample points add to one and the distributive law for numbers,

$$E(c) = \sum_{s\in S} cP(\{s\}) = c\sum_{s\in S} P(\{s\}) = c.$$

Multiplication: Using the distributive law again

$$E(cX) = \sum_{s\in S} cX(s)P(\{s\}) = c\sum_{s\in S} X(s)P(\{s\}) = cE(X).$$

Demonstration

Additivity: Using the definition of expectation in (8), and the distributive, commutative and associative laws for numbers

$$\begin{aligned} E(X+Y) &= \sum_{s\in S}(X(s)+Y(s))P(\{s\}) \\ &= \sum_{s\in S}X(s)P(\{s\}) + \sum_{s\in S}Y(s)P(\{s\}) \\ &= E(X)+E(Y). \end{aligned}$$

Linear property follows from induction using the case n=2 and both the addition and multiplication properties in order:

$$\begin{aligned} E(c_1X_1+c_2X_2) &= E(c_1X_1)+E(c_2X_2) \\ &= c_1E(X_1)+c_2E(X_2). \end{aligned}$$

2.7 Application to Variance

Application of Properties of Expectation to Variance

Using the definition of variance in (13), expanding the quadratic for each sample outcome, recalling that $\mu = E(X)$ and using the Linearity property of expectation gives (18)

$$\begin{aligned} Var(X) &= E\big[(X-\mu)^2\big] \\ &= E\big[X^2-2\mu X+\mu^2\big] \\ &= E(X^2)-2\mu E(X)+\mu^2 \\ &= E(X^2)-\big[E(X)\big]^2. \end{aligned} \tag{19}$$

2.8 Example: Variance of Triangular PMF

Example: Variance of Triangular PMF

Find the variance of a random variable which has probability mass function f given by

$$f(x) = \begin{cases} 1/9 & x=1 \text{ or } x=5 \\ 2/9 & x=2 \text{ or } x=4 \\ 3/9 & x=3 \end{cases}$$

Example: Triangular PMF

Figure 12 shows the pmf.

Solution: Variance of Triangular PMF

Mean:

$$E(X) = (1+5)\times\frac{1}{9}+(2+4)\times\frac{2}{9}+3\times\frac{3}{9} = 3.$$

Intuitive that mean is 3 because the PMF is symmetric about 3

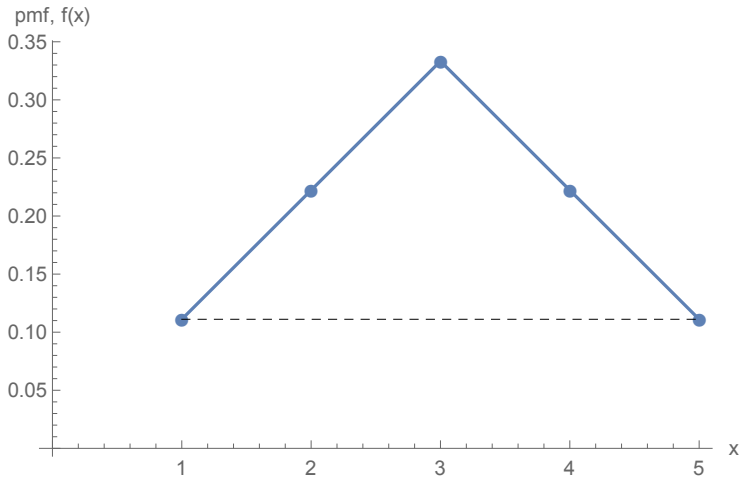

Figure 12: Find Mean and Variance for PMF

Mean Square:

$$E(X^2) = (1+5^2) \times \frac{1}{9} + (2^2+4^2) \times \frac{2}{9} + 3^2 \times \frac{3}{9} = 10\frac{1}{3}\,.$$

Variance:

$$Var(X) = E(X^2) - \left[E(X)\right]^2 = 10\frac{1}{3} - 9 = 1\frac{1}{3}\,.$$

2.9 Example: Minimum Squared Deviation

Example: Mean Minimises Squared Deviations
Find the value of the number b which minimises $E\left[(X-b)^2\right]$.

Solution: Mean Minimises Squared Deviations

Calculus is not needed — text Example 2.2-4 does this example but uses calculus.

Method: Guess and verify by the "put in what you want and then make the necessary correction" method.

μ? Using the three properties of expectation in an extension of the argument used to find

$$\begin{aligned} E\left[(X-b)^2\right] &= E\left[[(X-\mu)-(b-\mu)]^2\right] \\ &= E\left[(X-\mu)^2\right] - 2(b-\mu)E(X-\mu) + (b-\mu)^2 \\ &= E\left[(X-\mu)^2\right] + (b-\mu)^2, \end{aligned}$$

Since the middle, cross-product term, is 0 : $\mu = E(X)$!

RHS is $\geq Var(X)$ since the first term is the variance and the second is ≥ 0 and 0 if, and only if, $b = \mu$.

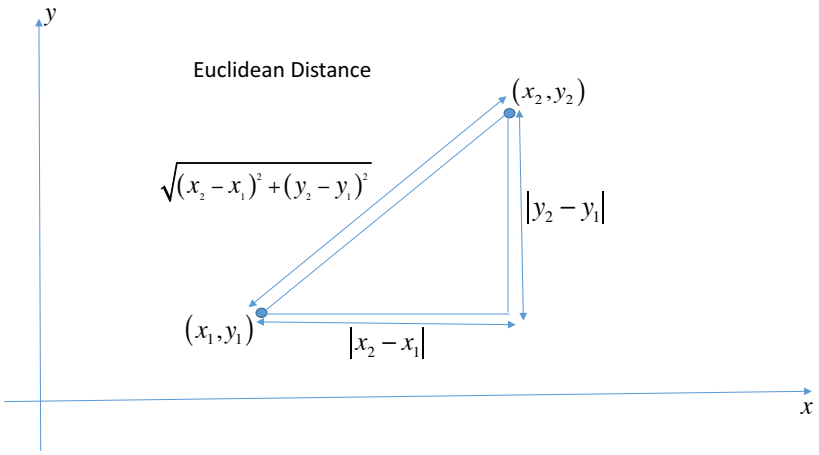

Figure 13: Pythagoras Theorem gives Euclidean Distance

Comment: Mean Minimises Squared Deviations

For data, descriptive statistics likes to make a numerical summary of all the values: data value = summary + residual

For data $x_1, x_2, \ldots, x_n$ and any number b as a summary, this becomes $x_i = b + (x_i - b)$.

Criterion: b should be chosen to minimise the residuals in total, for example to minimise the distance of the residuals from the origin.

Euclidean Distance of the residuals from the origin is

$$\sqrt{\text{sum of squares of residuals}} = \sqrt{\sum_{i=1}^{n} (x_i - b)^2}.$$

Euclidean Distance Reminder

Figure 13 shows Pythogoras' Theorem.

Comment: Mean Minimises Squared Deviations

Particular case of the example, is the random variable, X, that chooses one of the x_i at random.

The expected value of $(X - b)^2$, is related to the Euclidean distance by

$$\text{Euclidean distance} = \sqrt{n * E(X - b)^2}.$$

Hence Euclidean distance of the residuals from the origin is minimised by choosing $b = \bar{x}$.

Analogy for a general random variable is that $b = \mu = E(X)$ minimises a weighted Euclidean distance of the residuals from the origin with the weights proportional to the probabilities of the values of X.

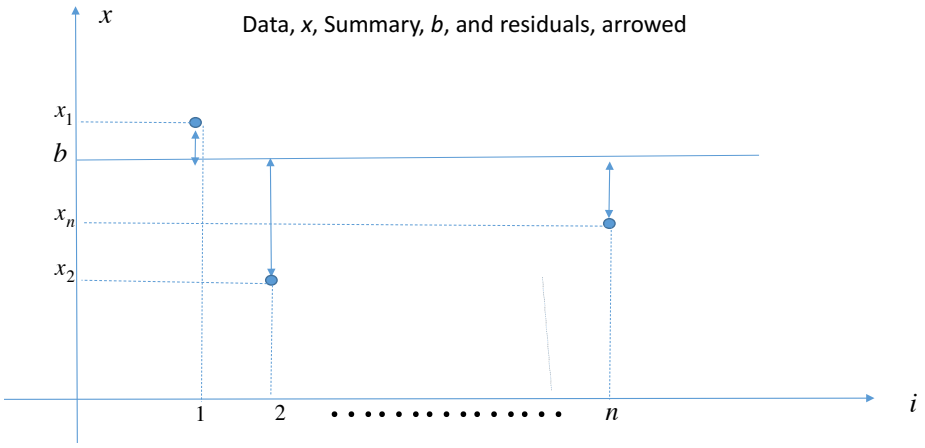

Figure 14: Data, Summary and Residuals

Data = Summary + Residuals

Figure 14 shows the data, residuals and summary value.

Standard Deviation (Ch 2.3) — Definition

Euclidean Distance takes the square root so that it is in the same units as the measurements.

Standard Deviation of a random variable is the square root of $Var(X)$, often denoted σ.

2.10 Example: Sampling with and without replacement

Example: Mean Number with a Disease

Find the mean number of people with a particular disease in a random sample taken from a population with or without replacement. Assume that the sample size is n and that the population has a proportion p of people with the disease.

Solution: Mean Number with a Disease

Definition Let X_i $(i = 1, 2, \ldots, n)$ be 1 if the ith person sampled has the disease and 0 otherwise.

Let X be the rv which records the number of persons in the sample who have the disease so

$$X = \sum_{i=1}^{n} X_i.$$

Linearity of expectation gives

$$E(X) = \sum_{i=1}^{n} E(X_i).$$

With replacement makes the sample outcomes — diseased or not — the results of Bernoulli trials with constant probability p.

Note that for *any* random variable, Z, that only takes the values 0 and 1, $E(Z) = 1 \times P(Z = 0) + 0 \times P(Z = 0) = P(Z = 1)$.

Solution Mean Number with a Disease Ctd

So for sampling with replacement is $E(X) = nE(X_1) = np$.

Without replacement: perhaps not so clear, but true, that $E(X_i) = p, i = 1, 2, \ldots, n$, so that the answer for the mean number in the sample with the disease is still np.

The reason is that if the population size is t, we can imagine random orderings of *all* members of the population with equally likely outcomes for each random ordering.

Without replacement sample can then be taken as the first n members of the random *ordering* of the *whole* population.

Because each sample of the first n has the same number of orderings of the whole population that produce it, the first n in the ordering still produces equally likely samples.

Solution Mean Number with a Disease Ctd 2

Then using the same definition for X_i $(i = 1, 2, \ldots, n)$,

$$\begin{aligned} P(X_i = 1) &= \frac{\text{number of outcomes in } [X_i = 1]}{\text{total number of outcomes}} \\ &= \frac{b \times (t-1)!}{t!} \\ &= \frac{b}{t} = p, \end{aligned}$$

since whichever of the b people with the disease are chosen in position i in the random ordering, there are $(t-1)!$ orderings for the other $t-1$ positions in the ordering.

Thus, as already claimed, $E(X) = nE(X_1) = np$ for sampling with and without replacement.

We'll return to the variance for sampling with replacement after discussing Moment Generating Functions and Independent Random Variables.

3 Moment Generating Functions — 2.3

3.1 Definition

Definition of Moment Generating Function

Suppose X is a random variable.

Then the *Moment Generating Function* $M(t)$ for a number t is defined as

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= \sum_{x \in range(X)} e^{tx} P(X = x) \\ &= \sum_{x \in range(X)} e^{tx} f(x), \end{aligned}$$

where $f(x)$ is the probability mass function of X.

May not exist if X has infiintely many values but does exist at $t = 0$ and usually does for $-h < t < h$ for some $h > 0$.

MGFs are a convenient mathematical tool especially for sums of independent random variables (for example, the Binomial rv).

3.2 Example: Moment Generating Function from PMF

Example: MGF of Triangular Random Variable

Find the moment generating function, and its derivative, for the random variable X whose probability mass function is the triangular one:

$$f(x) = \begin{cases} 1/9 & x = 1 \text{ or } x = 5 \\ 2/9 & x = 2 \text{ or } x = 4 \\ 3/9 & x = 3 \end{cases}$$

What is the value of the derivative at 0? Interpret for X.

Solution: MGF of Triangular Random Variable

MGF

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= 1/9 e^t + 2/9 e^{2t} + 3/9 e^{3t} + 2/9 e^{4t} + 1/9 e^{5t}. \end{aligned}$$

Derivative $M'(t) = 1/9 e^t + 2 \times 2/9 e^{2t} + 3 \times 3/9 e^{3t} + 4 \times 2/9 e^{4t} + 5 \times 1/9 e^{5t}$.

So $M'(0) = 1/9 + 2 \times 2/9 + 3 \times 3/9 + 4 \times 2/9 + 5 \times 1/9 = 3$.

Interpretation: a complicated way to find $E(X)$!

3.3 Example: Moments from MGF

Why the name?

MGF M of rv X satisfies $M(t) = E(e^{tX})$. Differentiating with respect to t, under the sum which is the expectation, gives

$$M'(t) = E(Xe^{tX}).$$

Putting $t = 0$ gives

$$E(X) = M'(0). \tag{20}$$

Differentiating again

Gives

$$E(X^2) = M''(0). \tag{21}$$

So

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ &= M''(0) - [M'(0)]^2. \end{aligned} \tag{22}$$

In general

Differentiating k times $k = 0, 1, 2, \ldots$

$$E(X^k) = M^{(k)}(0), \tag{23}$$

where the brackets (k) in the numerator indicates the kth derivative.

Note

$$1 = E(X^0) = M(0).$$

Example: Finding Moments from MGF

Find the mean and variance of a random variable whose moment generating function is $M(t) = (q + pe^t)^n$ where $0 < p < 1$, $q = 1 - p$ and n is a positive integer.

Solution: Finding Moments from MGF

Derivative

$$M'(t) = pe^t \times n(q + pe^t)^{n-1}.$$

Mean $E(X) = M'(0) = np$.

2nd Deriv

$$M''(t) = pe^t \times n(q + pe^t)^{n-1} + p^2e^{2t} \times n(n-1)(q + pe^t)^{n-2}.$$

So $Var(X) = M''(0) - [M'(0)]^2 = np + n(n-1)p^2 - (np)^2 = np(1-p)$.

3.4 PMF from MGF

PMF from MGF

In general if the MGF exists $-h < t < h, h > 0$, then the PMF can be found from the MGF.

If it is known that there are (say) 5 values in the range of the rv, then evaluating the MGF at 5 different points will give 5 simultaneous equations in the 5 unknown probabilities that can be solved.

More complicated for other random variables, for example if there are infinitely many possible values.

However, the mgf is always good for finding expectations of the random variable to all powers.

Since these can be found by repeated differentiation and substition of $t = 0$.

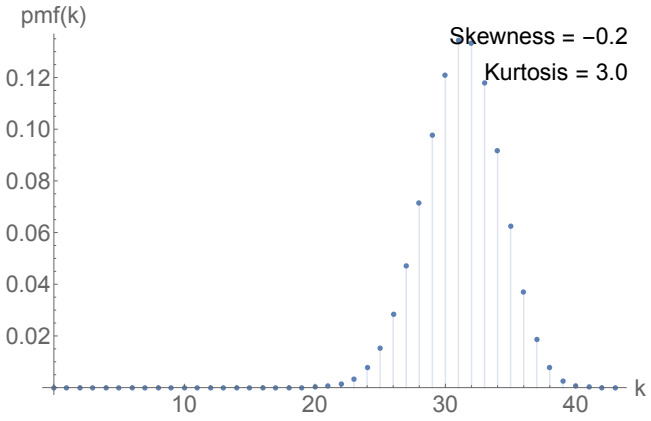

Figure 15: One example: Mathematica Manipulate in Next Week's Lab

Moments Tell About Shape

Mean gives the central value for the pmf.

Variance gives the extent of spread of the pmf around the mean.

Standardised random variable is $\frac{X-\mu}{\sigma}$ where μ is the mean and $\sigma = \sqrt{Var(X)}$ is the standard deviation.

Skewness is the third moment for the standardised random variable — pmf skewed to right or left

So skewness is $E\left[\left[\frac{X-\mu}{\sigma}\right]^3\right]$.

Kurtosis is the fourth moment for the standardised random variable — pmf flat or peaked.

So kurtosis is $E\left[\left[\frac{X-\mu}{\sigma}\right]^4\right] \geq 0$.

Example: Finding Moments About PMF
Figure 15 is from next week's lab.

4 Independent Random Variabless — 4.1

4.1 Indep't RVs — Var, MGF of Sum

Definition and Main Property, Variance and MGF

Definition Random variables X and Y are *independent*, if, for any sets of values B_1, B_2, $[X \in B_1]$ and $[Y \in B_2]$ are independent events, that is

$$P(X \in B_1 \cap Y \in B_2) = P(X \in B_1)P(Y \in B_2)$$

Extension to *independent* random variables $X_1, X_2, \ldots, X_n$ requires for any sets of values $B_1, B_2, \ldots, B_n$, that the events $[X_i \in B_i]$, $i = 1, 2, \ldots, n$ are independent

Main property (which will be shown in Module 4)

$$E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n) \tag{24}$$

Useful If $X_1, X_2, \ldots, X_n$ are *independent*, so are $f_1(X_1), f_2(X_2), \ldots, f_n(X_n)$ for any functions $f_1, f_2, \ldots f_n$

Application to Variance of a Sum

If X and Y are independent random variables, then

$$Var(X + Y) = Var(X) + Var(Y) \tag{25}$$

Demonstration

By linearity of expectation, the calculation formula for variance and the main property for *independent* random variables

$$\begin{aligned} Var(X + Y) &= E[(X + Y)^2] - (E(X) + E(Y)))^2 \\ &= E[X^2 + Y^2 + 2XY] - (E(X) + E(Y))^2 \\ &= Var(X) + Var(Y) + 2E(XY) - 2E(X)E(Y) \\ &= Var(X) + Var(Y) \end{aligned} \tag{26}$$

just like Pythagoras' theorem.

Variance of sum for *non-independent* random variables

Can be calculated using using equation (26) on the last slide since independence only enters in using the Main Property (equation (24)) for independent random variables

This will be discussed further in Module 4.

The variance for the number of successes in sampling *without* replacement is an example — see later slides for pictures of the pmf in this case.

4.2 MFG of a Sum

Application to MGF of a Sum

If X and Y are *independent* random variables, then so are e^{tX} and e^{tY}, so

$$\begin{aligned} M_{X+Y}(t) &= E(e^{t(X+Y)}) \\ &= E(e^{tX} e^{tY}) \\ &= E(e^{tX})E(e^{tY}) \\ &= M_X(t)M_Y(t) \end{aligned} \tag{27}$$

4.3 Properties of Variance

Analogy to Expectation

Constant If a is a constant, then $Var(a) = 0$

Multiplication If a is a constant and X is a random variable, then

$$Var(aX) = a^2 Var(X)$$

Addition If a, b are constants and X, Y are *independent* random variables, then

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Linear Comb If $a_1, a_2, \ldots, a_n$ are constants and $X_1, X_2, \ldots, X_n$ are *independent* random variables, then

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i)$$

Sample Mean of Independent RVs — All Same Distribution

So If $X_1, X_2, \ldots, X_n$ are *independent* random variables *each with the same pmf*, then

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\ &= \sum_{i=1}^{n} n^{-2} Var(X_i) \\ &= n^{-2} \times n \times Var(X_1) \\ &= \frac{Var(X_1)}{n} \end{aligned} \tag{28}$$

As n gets large, the variance of the sample mean gets small, i.e., most values are concentrated around the population mean, $\mu = E(X)$

4.4 Properties of Sample Mean

Any RVs — All Same Distribution

Using linearity of expectation (as in the example for the mean number having a disease)

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\ &= n^{-1} \times n \times E(X_1) \\ &= E(X_1) \end{aligned} \tag{29}$$

Summary: Moments of Sample Mean

If the mean and standard deviation in the population are μ and $\sigma = \sqrt{Variance}$, then the sample mean $\bar{X}$ of n *independent and identically distributed random variables* from this population has expectation μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

This is why the amount of information collected increases generally with the *square root* of the sample size

In Module 5 we'll study this relationship with sample size more

These connections are basic tools in practical statistical data science

4.5 Sample Proportion

Example: Sample Proportion

Find the mean, variance and standard deviation of the random variable which gives the sample proportion in a random sample taken intending to vote Liberal at the next election if the sample is taken with replacement and has size 2000. Assume the population proportion intending to vote Liberal is (an unknown number) p. Interpret. If the sample is taken without replacement, when will these answers give an accurate guide?

Solution: Sample Proportion

The sample proportion is $\bar{X}$ for the random variables $X_1, X_2, \ldots, X_{2000}$ with $X_i = 1$ if the ith person in the sample intends to vote Liberal and $X_i = 0$ otherwise.

So using equation (29), the expectation or mean of the random variable which gives the sample proportion is p, the population proportion intending to vote Liberal, since $E(X_1) = 1 \times p + 0 \times (1 - p)$.

Also this is true with or without replacement in the sample

Interpretation: Repeated observations of the sample proportion — for example, repeated opinion polls (provided the population proportion of Liberal intending voters is not changing) — will centre around p, the population proportion.

Solution: Sample Proportion

Using (28) for random samples *with replacement*, the sample proportion random variable has variance $\frac{p(1-p)}{2000}$ and standard deviation $\sqrt{\frac{p(1-p)}{2000}}$.

Interpretation is that the spread of the distribution for the sample mean is small in a sample of 2000, since the standard deviation is maximised when $p = \frac{1}{2}$ and is then 0.011.

When the sample is taken without replacement, the expectation is exact.

Further the variance and standard deviation will be approximately right since the population size of enrolled voters is about 16 million and this is *much* larger than the sample size.

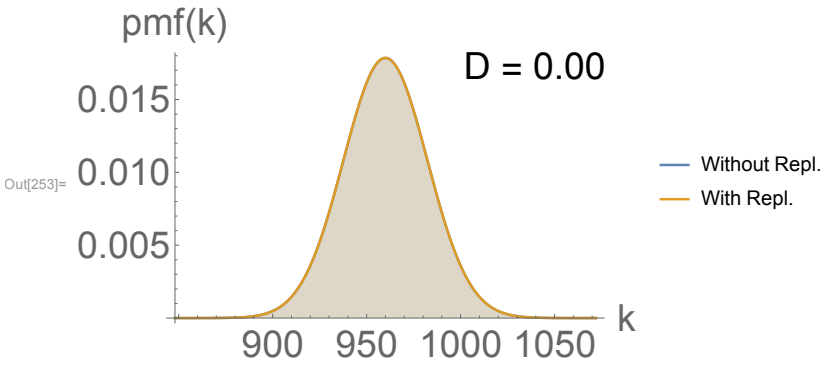

Figure 16: Electorate=16m, p= 0.48 *NewsPoll* 20 *March* 2017

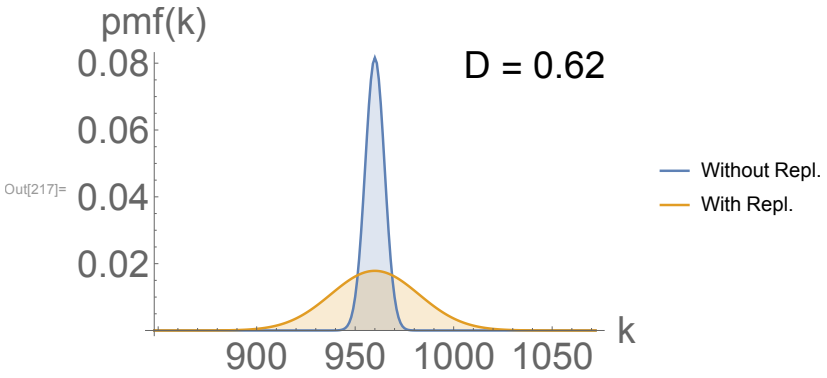

Figure 17: Pop Size=2,100, p= 0.48 *NewsPoll* 20 *March* 2017

PMF — With & Without Replacement

Definition *D* in the pictures of the pmf's with and without replacement is the maximum difference in *any* probability computed with or without replacement

Plot shows no difference for any probability in sampling with or without replacement for a population as large as 16 million, the size of the Australian Electorate, when the sample size is 2000

On the other hand if the total size were just 2100 then there is a big difference in probabilities with and without replacement, as Figure 17 shows.

Discussion — With & Without Replacement

Shape very similar in both population sizes, 16 million and 2,100

Mean the same

Variance much smaller for 2100

What is the variance for population size 2100? See Module 4 for variance of hypergeometric

Figures 18 & 19 show the PMFs with and without replacement for pop'n sizes 3,100 and 10,100

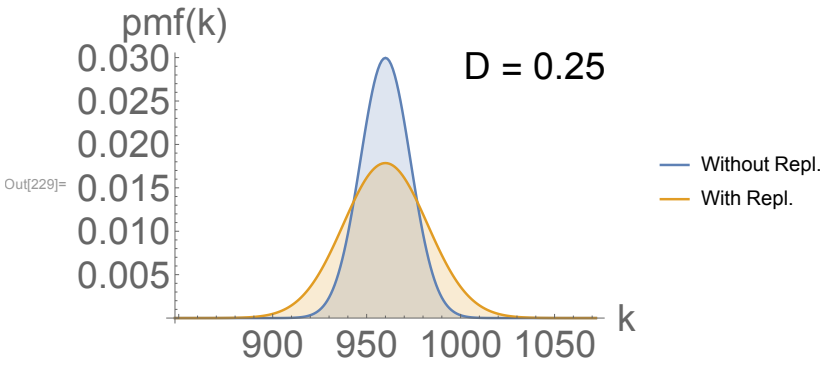

Figure 18: Pop Size=3,100, p= 0.48 *NewsPoll* 20 *March* 2017

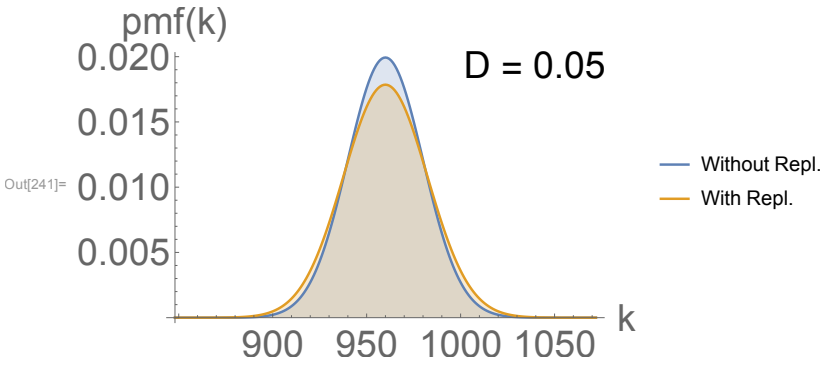

Figure 19: Pop Size=10,100, p= 0.48 *NewsPoll* 20 *March* 2017

PMF Shape Only

Standardized random variable, Z, is defined as

$$\frac{X-\mu}{\sigma}$$

for any random variable X with $E(X) = \mu$ and $Var(X) = \sigma^2$

Z takes value 0 if $X = \mu$, value $\pm i, i = 1, 2, \ldots$ if $X = \mu \pm i\sigma$, so *standardizing* changes units for X to centre at the mean and scale with the standard deviation

Reason for standardizing will become clearer in Module 5, but showing the pmf for the standardized random variable removes any differences between random variables in both mean and variance

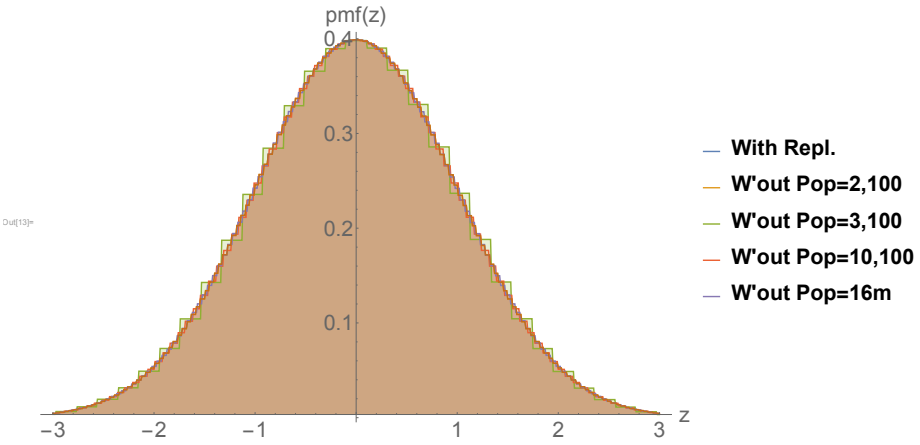

Figure 20: Scaled pmf for p=0.48, with and w'out replacement, Pop Size varying

Because the units on the z axis increase by the reciprocal of the standard deviation, to preserve the probabilities adding to one, it is necessary to scale the pmf by the stand. dev.

Next figure shows the pmf for the *standardized* sample number with replacement, and without replacement for population sizes 2,100, 3,100, 10,100 and 16,000,000

5 Binomial Distribution — 2.4

5.1 Recall PMF from Module 2_1

Recall Bernoulli Trials from Module 2_1

Bernoulli Trials a number, n, of *independent trials*.

Two Outcomes: each trial has one of two possible outcomes *success* or *failure* denoted S and F.

Events $A_i, i = 1, \ldots, n$ where A_i occurs if there is *success* on the ith trial.

Binomial RV X is the number of trials that result in success.

Component RVs For $i = 1, 2, \ldots, n$, let X_i be the rv which is 1 on A_i and 0 on A_i^c, so that X_i is 1 for a success on trial i and 0 for a failure on trial i.

Now The *number* of successes overall is X and this is the sum of the number of successes on each trial so

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i. \tag{30}$$

PMF of Binomial RV

Probability of Success Let p be the constant probability of success on each trial

Earlier we showed equation 2 that for $x = 0, 1, 2, \ldots, n$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

5.2 Binomial Mean and Variance

Mean and Variance of Binomial RV

From the representation (30) and the properties of the Sample Mean of Independent Random Variables with the same Distribution

$$E(X) = np$$

And

$$Var(X) = np(1-p)$$

5.3 Sampling with Replacement

Sampling with Replacement
is a special case and the number in the sample has Binomial distribution with n equal to the sample size and p equal to the population proportion of "successes". The number with a disease and sample proportion voting Liberal were examples of the Binomial distribution.

5.4 MGF for Binomial

MGF for Binomial

Bernoulli random variable X_1 has MGF M_1 given by

$$\begin{aligned} M_1(t) &= E(e^{tX_1}) \\ &= e^{t\times 1} P(X_1 = 1) + e^{t\times 0} P(X_1 = 0) \\ &= (pe^t + q) \end{aligned}$$

where $q = 1 - p$

MGF for sum of *independent* rv's is product of MGF's

So

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= M_1(t) M_2(t) \cdots M_n(t) \\ &= (pe^t + q)^n \end{aligned}$$

If an MGF is this form, then the pmf is Binomial (because it is defined for all t).

6 Negative Binomial Distribution — 2.5

6.1 PMF for Time till Success

Time till First Success in Bernoulli Trials

Suppose $X_1, X_2, \ldots$ are random variables which record 1 on successes and 0 on failures in an indefinitely long sequence of Bernoulli trials

So the X random variables have $P(X_i) = E(X_i) = p$ for each $i = 1, 2, \ldots$ and they are independent

Let T_1 be the random variable which counts the number of trials until a success occurs

Note on Time/Trial Number: trials number is often referred to as "time" which may, or may not, correspond to actual time depending on the context

Then the probability mass function for T is given by

$$\begin{aligned} P(T_1 = k) &= P(X_1 = 0, \ldots, X_{k-1} = 0, X_i = k) \\ &= P(F \cdots FS) \\ &= (1-p)^{k-1}p, \quad k = 1, 2, \ldots \end{aligned} \tag{31}$$

Time till rth Success in Bernoulli Trials

Argument is similar for $T_r, r = 2, \ldots$ which is the number of trials till the rth success.

Any pattern which has $T_r = k$ has exactly r successes in the first k trials with the rest failures. Also, trial k must be a success on the event $T_r = k$. Apart from a success on trial k, the trials for the other $r-1$ successes in the first $k-1$ trials can be chosen freely.

There are $\binom{k-1}{r-1}$ ways to choose the trial numbers for the first $r-1$ successes since they must occur in the first $k-1$ trials.

So

$$P(T_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \quad k = r, r+1, \ldots.$$

Geometric and Negative Binomial

Geometric is the name for the distribution for the time to the first succcess in Bernoulli trials.

There is just one number p, the probability of a success on a trial, that determines the pmf.

Negative Binomial is the name for the distribution of the time to the rth success.

Negative Binomial has two numbers, r and p which determine the pmf.

Confusingly your textbook and software (both Mathematica and R) disagree on the definition.

Mathematica uses the number of *failures* before the first, second, ... success and uses the parameter n rather than r for the number of successes.

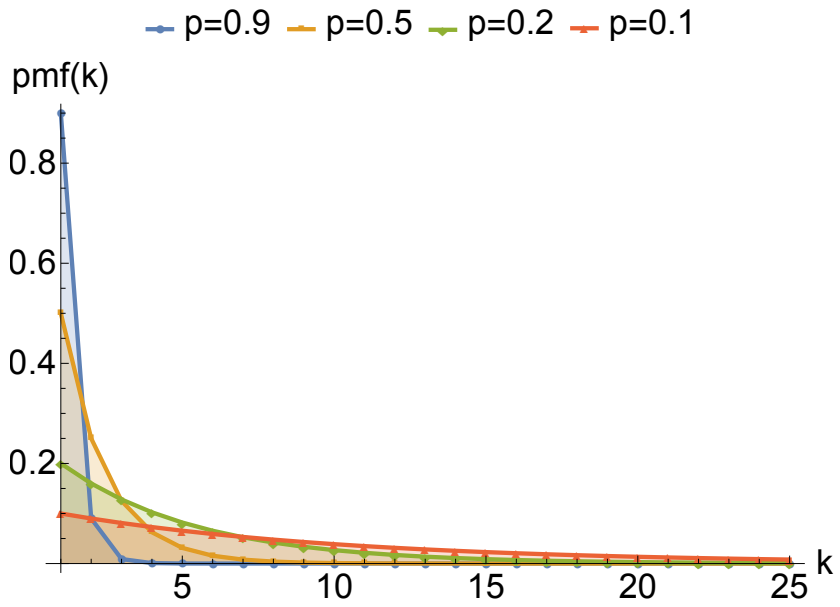

Figure 21: Prob Mass Functions for Geometric

6.2 Examples of PMF for Negative Binomial Distribution

To see shapes

Need to standardize as with Binomial and Hypergeometric pictures. Scale to rv $Z = \frac{X - \mu}{\sigma}$ and scale the pmf of this so the area under the curve is 1. Figures 26, 27 and 28 show the shapes by displaying the scaled pmf for Z.

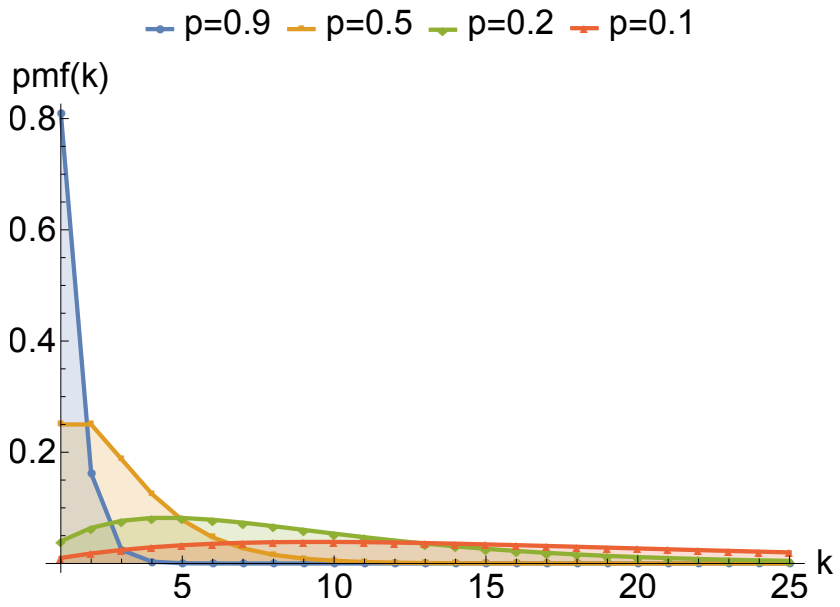

Figure 22: PMFs for Trial Number of 2nd Success in Bernoulli Trials

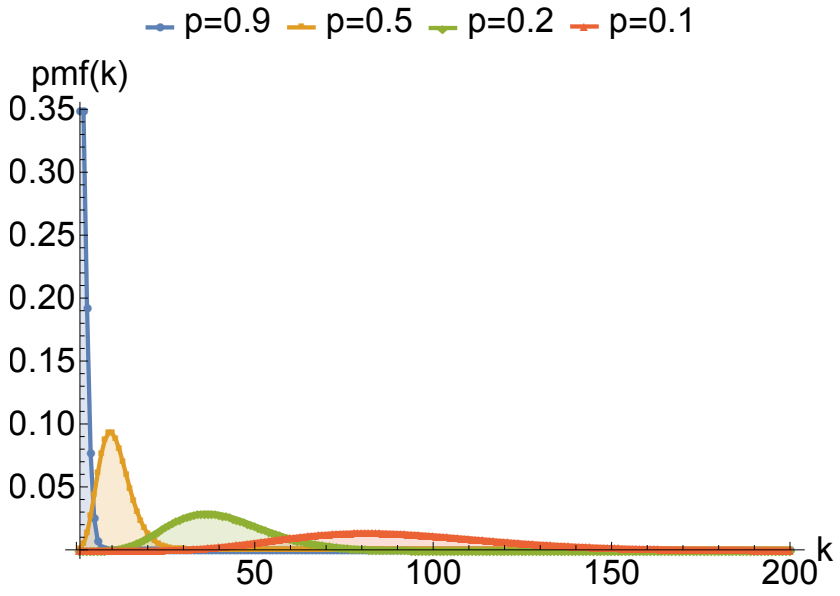

Figure 23: PMFs for Trial Number of 10th Success in Bernoulli Trials

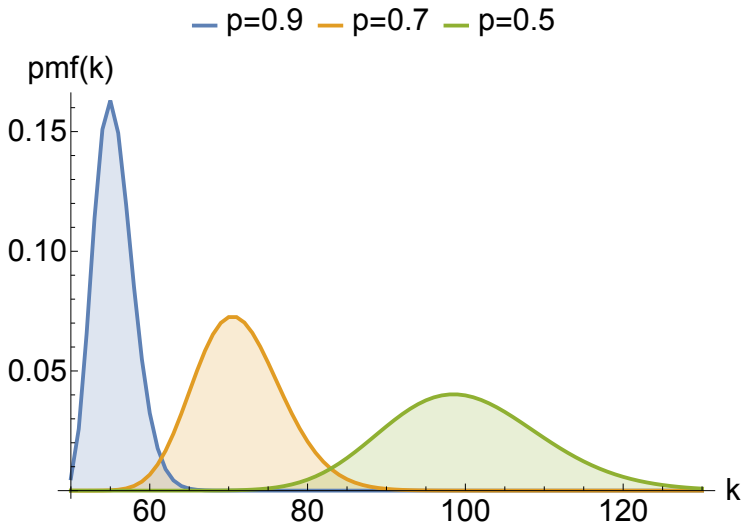

Figure 24: PMFs for Trial Number of 50th Success in Bernoulli Trials, High Probs

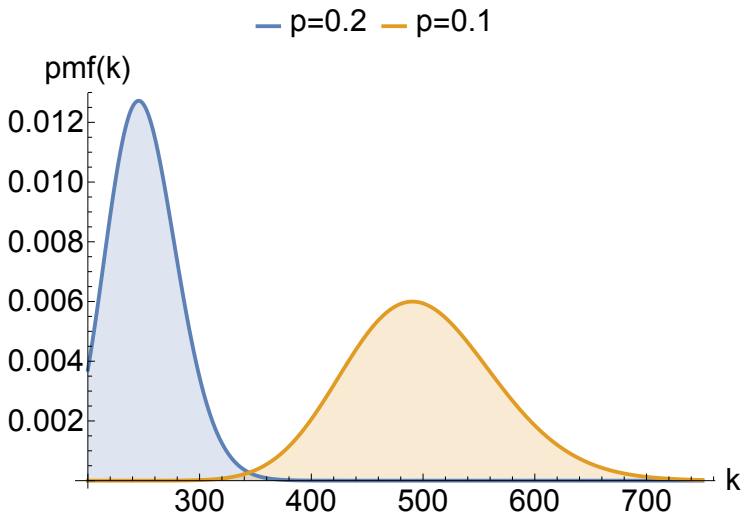

Figure 25: PMFs for Trial Number of 50th Success in Bernoulli Trials, Low Probs

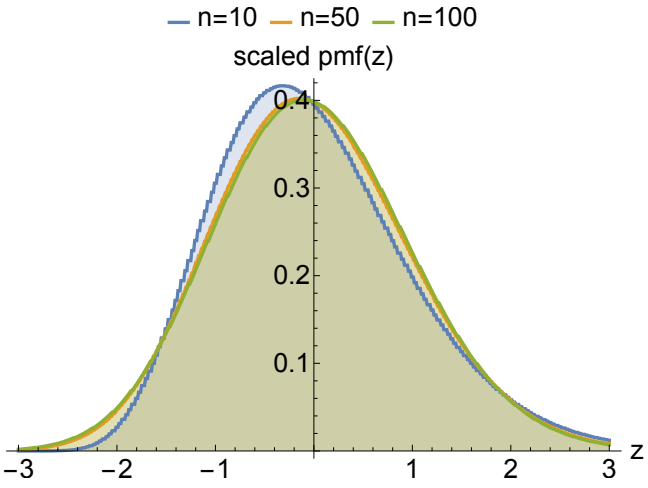

Figure 26: Standardized Time till rth success, p=0.1, r=10,50,100

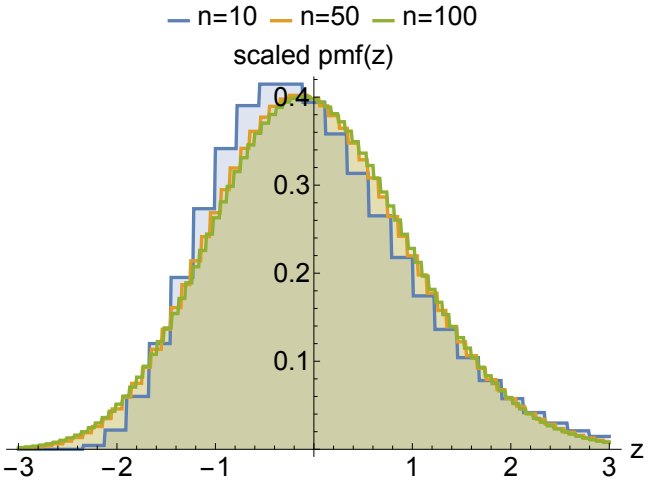

Figure 27: Standardized Time till rth success, p=0.5, r=10,50,100

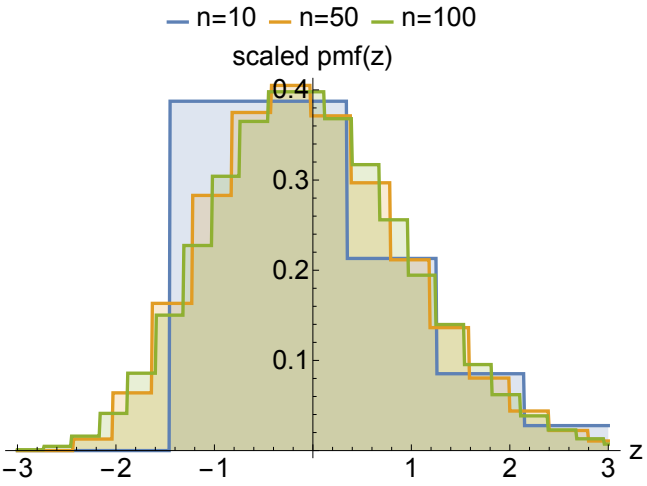

Figure 28: Standardized Time till rth success, p=0.9, r=10,50,100

Conclusion

For p=0.1 and p=0.5, as r gets larger see same shape as for Binomial and Hypergeometric.

For p=0.9, the successes come more quickly and the shape is not so clear.

Note: Skewness is evident for $r = 10$.

Also there is some probability for $|Z| > 3$.

Why the name Negative Binomial?

Binomial formula for positive powers, writing $q = 1 - p$, demonstrates that the binomial probabilities add to one :

$$1 = (p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^x q^{n-k}. \tag{32}$$

Binomial formula for *negative powers* says: for $-1 < w < 1$

$$(1-w)^{-r} = \sum_{k=r}^{\infty} \binom{k-1}{r-1} w^{k-r}. \tag{33}$$

So the sum of the pmf of the Negative Binomial is, substituting $q = w$,

$$\sum_{k=r}^{\infty} \binom{k-1}{r-1} p^r q^{x-r} = p^r (1-q)^{-r} = 1$$

.

6.3 MGF and Moments for Negative Binomial

Independence of Times Between Successes

Trial number of rth success is the sum of the times between the first $r > 1$ successes:

$$T_r = T_1 + (T_2 - T_1) + \cdots + (T_r - T_{r-1}) \tag{34}$$

And the random variables in the sum are *independent* and each distributed as the Geometric(p) distribution

Because building on the case $r = 2$ by induction, for $i, j \geq 1$, letting $q = 1 - p$, "iFS" mean i failures followed by a success, and using independence of the Bernoulli trials,

$$\begin{aligned} P(T_1 = k \cap T_2 - T_1 = j) &= P(T_1 = k)P(T_2 - T_1 = j | T_1 = k) \\ &= P((k-1)FS)P((j-1)FS|(k-1)FS) \\ &= q^{k-1}pq^{j-1}p \\ &= P(T_1 = k)P(T_1 = j) \end{aligned}$$

MGF for Geometric

Using (33) in the case $r = 1$ (the Geometric series)

$$\begin{aligned} M_1(t) &= E(e^{tT_1}) \\ &= \sum_{k=1}^{\infty} e^{tk} P(T_1 = k) \\ &= \sum_{k=1}^{\infty} e^{tk} q^{k-1} p \\ &= pe^t \sum_{k=1}^{\infty} (e^t q)^{k-1} \\ &= \frac{pe^t}{1 - qe^t} \end{aligned} \tag{35}$$

MGF for Negative Binomial

In general the MGF of a sum of independent random variables is product of MGF's and from the independence of the random variables in (34) :

$$\begin{aligned} M_r(t) &= E(e^{tT_r}) \\ &= M_1^r(t) \\ &= \left(\frac{pe^t}{1 - qe^t} \right)^r \end{aligned} \tag{36}$$

Moments for Negative Binomial

The moments for the Geometric rv T_1 can be found from the MGF:

$$E(T_1) = M_1'(0) = \frac{1}{p} \;\&$$
$$Var(T_1) = M_1''(0) - (E(T_1))^2 = \frac{1-p}{p^2} \tag{37}$$

The moments for the General Negative Binomial rv T_r then frollow because the random variables in (34) are independent so:

$$E(T_r) = rE(T_1) = \frac{r}{p} \;\&$$
$$Var(T_r) = rVar(T_1) = \frac{r(1-p)}{p^2} \tag{38}$$

Mathematica and R

Both use $T_r - r$, the number of *failures* before the rth success, rather than T_r. This affects the values of times and means, but not variances and standard deviations, since for any constant a and rv X with mean μ_X :

$$\begin{aligned} Var(X+a) &= E\big((X+a-\mu_{X+a})^2\big) \\ &= E\big((X+a-(\mu_X+a))^2\big) \\ &= Var(X) \end{aligned} \tag{39}$$

6.4 Example: Accidents in a Workplace

Example: Accidents in a Workplace

In a workplace, there is a probability of 0.05 that there is an accident in any month. The number of accidents in successive months are independent.

1. What is the chance that there are no accidents in a year?
2. What is the chance that the 4th accident occurs more than ten years from now?
3. What is the mean and standard deviation of the time in years to the
   (a) first
   (b) fourth

   accident?

Solution: Accidents in a Workplace

Let T_r be the time to the rth accident and X_t be the number of accidents in t years, so that T_r has Negative Binomial (r, 0.05) pmf and X_n has Binomial($12n$, 0.05) pmf. Hence

1. the chance that there are no accidents in a year is $P(T_1 > 12) = P(X_1 = 0) = (1-0.05)^{12} = 0.54036$.

2. the chance that the 4th accident occurs more than 10 years from now is

$$\begin{aligned} P(T_4 > 120) &= P(X_{10} \leq 3) \\ &= (1-0.05)^{120} + 120 \times (1-0.05)^{119} \times 0.05 \\ &+ \frac{120 \times 119}{2}(1-0.05)^{118} \times 0.05^2 \\ &+ \frac{120 \times 119 \times 118}{3 \times 2}(1-0.05)^{117} \times 0.05^3 \\ &= 0.131707. \end{aligned}$$

Solution: Accidents in a Workplace Ctd

3. the mean and standard deviation of the time *in years*, by linearity properties of expectation, the mean and standard deviation divided by 12. Hence

   (a) the mean and standard deviation of the time *in years* to the first accident is $E(T_1)/12 = \frac{1}{12\times 0.05} = 1\frac{2}{3}$ and $\frac{\sqrt{Var(T_1)}}{12} = \frac{\sqrt{1-0.05}}{12\times 0.05} = 1.62447$

   (b) the mean and standard deviation of the time in years to the fourth accident is $E(T_4)/12 = \frac{4}{12\times 0.05} = 6\frac{2}{3}$ and $\frac{\sqrt{Var(T_4)}}{12} = \frac{\sqrt{4\times(1-0.05)}}{12\times 0.05} = 3.24893$

7 Poisson Distribution — 2.6

7.1 Definition and Derivation

Poisson introduced as Binomial Approximation

If accidents were reviewed on a daily basis, the mean number of accidents per day might be expected to be about 1/30 of the mean number in a month, i.e., 0.05/30.

If X is the number of accidents in a time period subdivided into n parts with p the probability of an accident in each one of the parts, then for any $A \subseteq \{0, 1, 2, \ldots\}$

$$\begin{aligned} &\left| P(X \in A) - \sum_{k\in A} e^{-np}\frac{(np)^k}{k!} \right| \\ &\quad = \left| \sum_{k\in A\cap\{0,\ldots,n\}} \binom{n}{k} p^k (1-p)^{n-k} - \sum_{k\in A} e^{-np}\frac{(np)^k}{k!} \right| \\ &\quad < p. \end{aligned} \tag{40}$$

Proof requires advanced techniques called Stein's method.

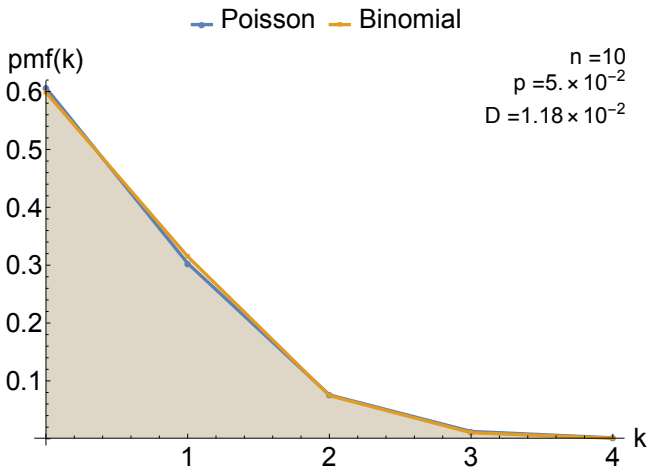

Figure 29: Binomial and Poisson Distributions, n=10, p=0.05

Definition of Poisson probabilities

The name for the probabilities that approximate Binomial is *Poisson*.

The approximation is close for small p regardless of n.

Definition: For any $\lambda > 0, k = 0, 1, \ldots$, the Poisson prob'y of k is

$$\frac{e^{-\lambda}\lambda^k}{k!}.$$

The number λ is called the *parameter* of the Poisson probabilities — n and p are the parameters of the Binomial distribution and np was the parameter of the approximate Poisson probabilities.

Figures 29,30 and 31 show how close are the Poisson and Binomial probabilities. The number D is the maximum difference between *any* Poisson and the corresponding Binomial probability.

Exponential Series shows the Poisson probs are a pmf

Taylor series for exponential function: for any x

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{41}$$

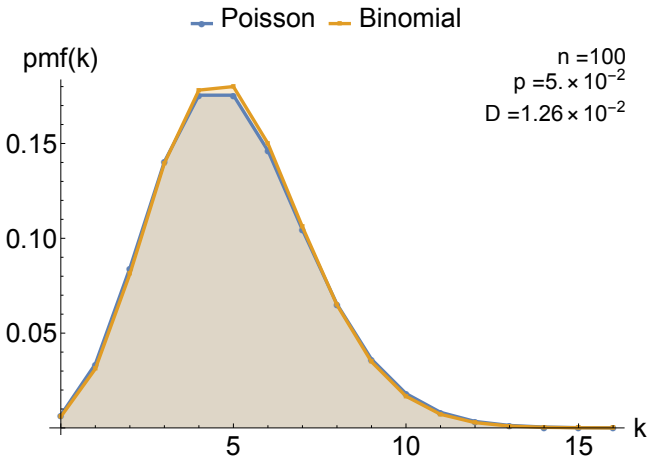

Figure 30: Binomial and Poisson Distributions, n=100, p=0.05

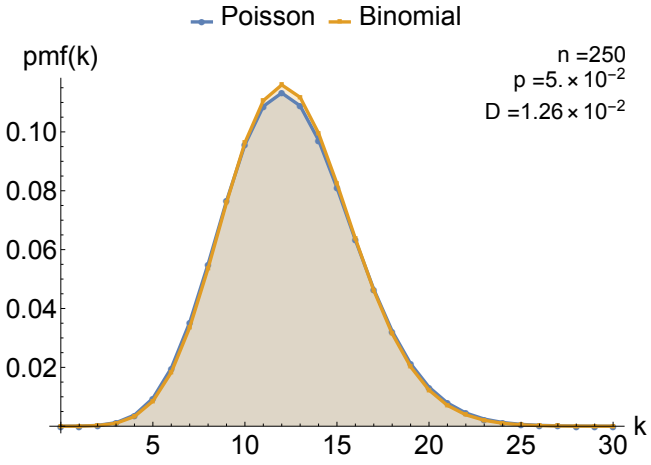

Figure 31: Binomial and Poisson Distributions, n=250, p=0.05

So the Poisson probabilities add to one: for any $\lambda > 0$ the sum of the Poisson probabilities is:

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1. \tag{42}$$

Poisson pmf, $f(k), k = 0, 1, 2, \ldots$ with parameter $\lambda > 0$ is defined as the kth entry of the series.

7.2 MGF and Moments

MGF and Mean of Poisson

If X has a Poisson pmf with parameter λ, then

$$\begin{aligned} M(t) &= E(e^{tX}) \\ &= \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)} \end{aligned} \tag{43}$$

So

$$E(X) = M'(0) = \left(\lambda e^t e^{\lambda(e^t-1)}\right)(0) = \lambda.$$

Variance of Poisson

MGF Second Deriv

$$M''(t) = \left(\lambda e^t\right)^2 e^{\lambda(e^t-1)} + \lambda e^t e^{\lambda(e^t-1)}$$

So

$$Var(X) = M''(0) - \lambda^2 = \lambda$$

7.3 Poisson Process and Example

Assumptions for Poisson process

Events occur in continuous time with no double occurrence of events at any *specific* time.

Counts of events in disjoint time intervals are independent random variables.

Equal mean number of events in time intervals of the same length.

Then an argument based on the Poisson approximation to Binomial shows that the number of events in any time t has *exactly* the Poisson distribution with parameter λt for some $\lambda > 0$.

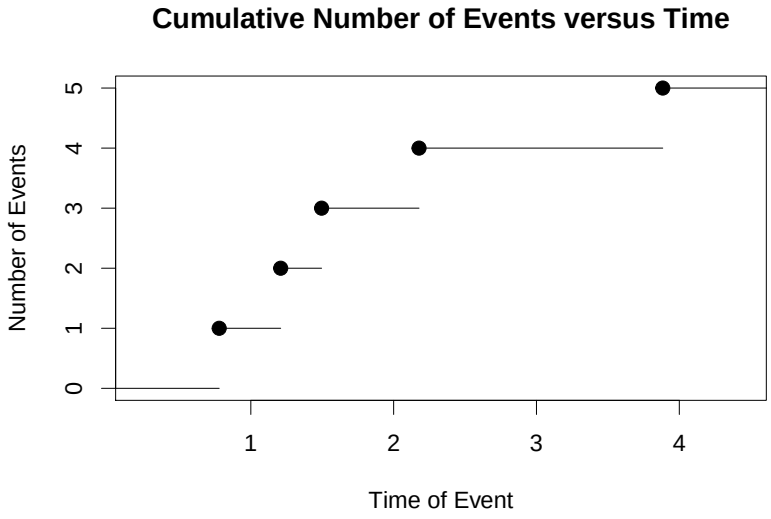

Figure 32: Assumptions require that the jumps are all size one

Poisson Process

The process of events is called a *Poisson process with rate* λ.

Interpretation of λ is the mean number of events in a unit time interval since this has an *exact* Poisson distribution with parameter $\lambda \times 1$.

If time is continuous, then the Binomial would only be an approximation to the Poisson — it ignores the possibility in our accident example that there would be *more* than one accident in a month.

Figure 32 shows an example of a Poisson process — the events occur at random in continuous time.

Example: Accidents in a Workplace with Poisson

Suppose that accidents in the workplace occur at a rate of 0.62 per year. Carry out the calculations for the first two parts of the Negative Binomial example with these assumptions. What is the mean and variance for the number of accidents per year? Comment on the answers.

Solution: Accidents in a Workplace with Poisson

Let T_r be the time to the rth accident and X_t be the number of accidents in t years, so that X_t has Poisson$(0.62t)$ pmf. Hence

1. the chance that there are no accidents in a year is $P(T_1 > 1) = P(X_1 = 0) = e^{-0.62} = 0.537944$,

2. noting that X_{10} has a Poisson distribution with parameter $\lambda = 10 \times 0.62 = 6.2$, the chance that the 4th accident occurs more than 10 years from now is

$$\begin{aligned} P(T_4 > 10) &= P(X_{10} \leq 3) \\ &= e^{-6.2} + \frac{e^{-6.2} \times 6.2}{1} \\ &+ \frac{e^{-6.2} \times 6.2^2}{2} + \frac{e^{-6.2} \times 6.2^3}{6} \\ &= 0.13423. \end{aligned}$$

Solution: Accidents in a Workplace with Poisson Ctd

Negative Binomial probabilities of 0.54036 and 0.131707 are very close, since they are based on the Binomial probabilities which are close.

Mean and variance for the number of events in a year are both the rate 0.62.

Mean for the Binomial(12,0.05) distribution is 0.6, close to 0.62.

Can match Binomial to Poisson by the mean or the probability of zero — see Lab this week.