Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

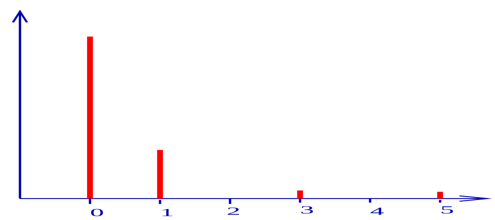
Lecture 7 (MWF) Sums of binary outcomes with an intro to hypothesis testing

Suhasini Subba Rao

Modelling the distribution of children

30 people in College Station are randomly sampled about the number of children they have (X number of children a person has and $X \in \{0, 1, 2, \ldots\}$). The data from this sample is summarized below.

No. of kids	0	1	2	3	4	5
frequency	22	6	0	1	0	1
probability	0.7	0.2	0	0.04	0	0.04



One may want to use this data to understand if the distribution of children is different to the national distribution etc.

Introduction

- To answer such questions we often rely on fitting models to the data.
- By fitting a model we answer several questions such as (a) check whether certain variables have an influence on an outcome (b) look for differences in distributions to name but a few.
- A common distribution for modelling the number of children where the possible outcomes are $\{0, 1, 2, 3, \ldots\}$ is a Poisson distribution.
- However, model fitting is not the main focus of this class.
- In this lecture we introduce the binomial distribution. We calculate the binomial probabilities in simple situations (by hand) and use software to calculate more complex probabilities.

- We use the binomial distribution as a devise for introducing the notion of a hypothesis test and as a motivation for the normal distribution.

- The binomial distribution is an important distribution in modelling. Modelling will form an important component of STAT652.

The binomial distribution

- This is an important distribution for modelling the distribution of categorical data.

- We often use it to test certain hypothesis. Eg. Whether more people are cured using a new drug treatment over an old treatment.

- Whether the proportion of people voting in elections now is different to the proportion in the past etc.

- It is used when several individuals are surveyed and the reply of each individual is binary. A binary variable is a categorical variable, where the number of choices is two. For example {Yes or No}, {Candidate A or Candidate B}.

- Typically, these variables are encrypted as $\{1 \text{ or } 0\}$. 1 or 0 are not probabilities, they are just a simple way to encode the reply.

- We assume that the response of each individual is independent of everyone elses response.

Example 1

Jack is a happy-go-lucky type of guy. He is so happy-go-lucky that he claims that he does not bother with revising his exam and simply guesses the answers. We want to see whether there is any truth in his claim.

- In a multiple choice exam (where there is an option of 5 questions) he has a 20% chance of getting the answer correct. If we try to write this formally we can let

$$\text{correct} \ = 1 \quad \text{wrong} \ = 0.$$

Let $X=$ either 1 or 0 depending on whether he gets it wrong or not..

P(He answer the question correctly)$=P(X=1)=0.2$

P(He answers a question incorrectly)$=P(X=0)=0.8.$

- Right or wrong are mutually exclusive events (Jack cannot be both right and wrong).

- Typically, we are not interested on the precise questions he answered correctly, but the total number of questions in the exam he answered correctly.

- If Jack selects each answer randomly, his score in his exam can take any value from zero to the highest number of marks in the exam.

- Let S_n denote the score out of n questions he did correctly. Then the set of all possible outcomes that S_n can take is $S_n = \{0, 1, \ldots, n\}$.

- We denote the probability he will score that he k the exam as $P(S_n = k)$ (for $0 \leq k \leq n$).

- If he guessed each question, then these probabilities follow a

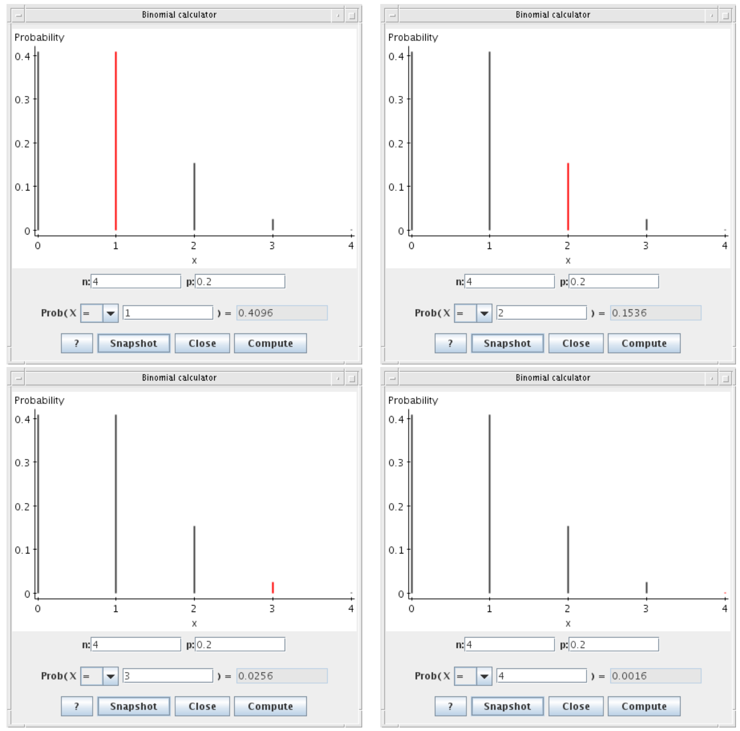
  Binomial distribution Bin$(n, p = 0.2)$, where n are the number of questions in the exam and $p = 0.2$ is the chance of him guessing each answer correctly.

Deriving the binomial distribution

- Using what we have learnt in Lecture 5 and 6, derive the distribution of $S_2 = X_1 + X_2$ (score when there are two questions in exam and $p = 0.2$, he guesses)

- Similarly, derive the distribution of the random variable $S_4 = X_1 + X_2 + X_3 + X_4$ (score in four questions and $p = 0.2$, he guesses).
  - It is clear that S_4 can take any of the values $\{0, 1, 2, 3, 4\}$.
  - Suppose Jack does 4 questions what is the probability he will get 1 answer correct?
    This can be written as $P(S_4 = 1)$.
  - Suppose Jack does 4 questions what is the probability he will get he will get 2 answers correct.
    That is $P(S_4 = 2)$?
  - Evaluate $P(S_4 = 0)$, $P(S_4 = 3)$ and $P(S_4 = 4)$.

Solution using software



Software plots the distribution (the probability of each possible outcome) and the probabilities.

The binomial distribution

This is a formal definition of the binomial distribution.

- Let X_i be the outcome of the ith trial (this is often called a Bernoulli trial). X_i can take the value $\{0, 1\}$ (eg. wrong or correct/yes or no). To these two outcomes we associate a probability $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$ (in the example above $P(X = 1) = 0.2$ and $P(X = 0) = 0.8$).

- Often

$$p = \text{proportion of "successes" in the population}$$

- We suppose that each trial is independent, that is $X_1, \ldots, X_n$ are independent random variables (for example, the chance Jack gets one answer correct is completely independent of the chance of Jack getting another correct).

- We may observe all the random sample $X_1, X_2, \ldots, X_n$. We are interested in the number of "successes" out of n, this is given by $S_n = X_1 + \ldots + X_n$.

- Since X_i is a random variable, then S_n is also a random variable which can take any one of the outcomes $\{0, 1, 2, \ldots, n\}$. Each outcome has a certain chance of occuring.

- This chance is given by the formula

$$P(S_n = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$$

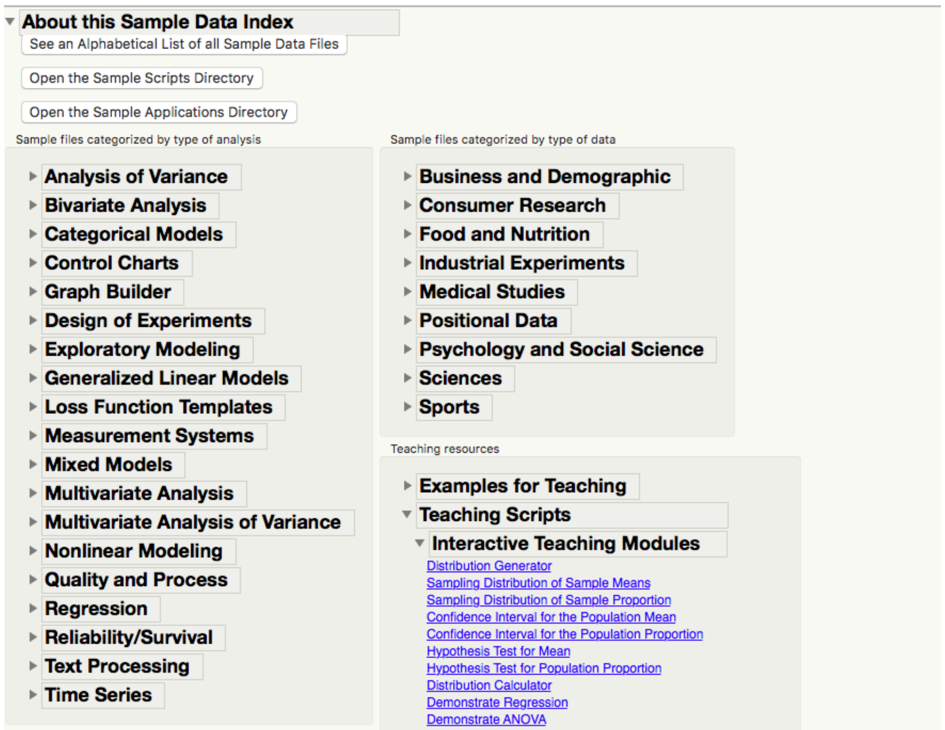
$n! = n \times (n-1) \times (n-2) \times \ldots \times 1$ $(0! = 1)$.

- Towards the end of the lecture we do some small calculations by hand. But we mainly rely on software to calculate the chance.

- **Notation** We often say that $S_n \sim Bin(n, p)$. To mean that the distribution of S_n is binomial, where the probability of a yes in each trial is p and number of trials n.

JMP: Calculating binomial and other probabilities

- To calculate binomial probabilities in JMP go to Help > Sample Data > Teaching Resources > Teaching Scripts > Interactive Teaching Modules. Select Distribution Calculator (which is highlighted in blue).



Example 2

Jack has taken his final exams. He boasts to his friends that he has been guessing all his answers.

He takes two multiple choice exams.

- In his Biology exam he scores 18 out of 30.
- In his Chemistry exam he scores 8 out of 30.

What do you think about his claims about simply randomly choosing the answer?

To answer this question we will utilize Statcrunch/JMP. But first we reformulate the question as a statistical test.

Formulating a hypothesis

- We formulate this question as a hypothesis test.

- There are two competing ideas (0) he guessed (A) he had some idea about the material.

- We are asking, based on his grades, if there is any evidence to "prove" that he knew the material (can we prove (A)).

- We state the two competing ideas as two competing hypothesis; the so called null hypothesis, denoted as H_0: is that he guessed.

  The competing hypothesis is usually called the alternative and denoted as H_A (or H_1): is that he knew some of the material.

- In terms of the binomial distribution $p = 0.2$ corresponds to the case he was guessing and $p > 0.2$ corresponds to the case that he knew some of the material.

- Formally, we rewrite the two "competing" hypotheses as

$$H_0 : p \leq 0.2 \text{ vs } H_A : p > 0.2.$$

- We can only "prove" H_A (prove the alternative hypothesis) by disapproving H_0 (disapprove the null hypothesis).

- We assess the validity of this claim (the validity of the null) by calculating the chance of obtaining the score he got or even better under the assumption his claim is true.

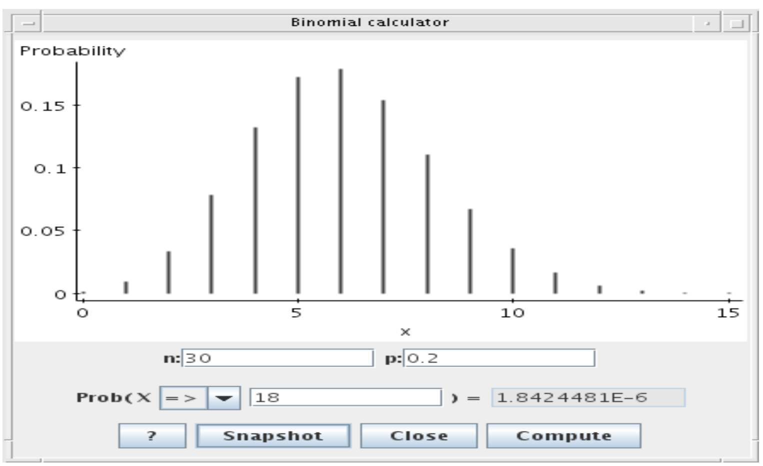
- Definition The probability of scoring 18 or greater under the claim he was guessing is called the **p-value**. The p-value is commonly used in statistical applications (though it can be problematic).

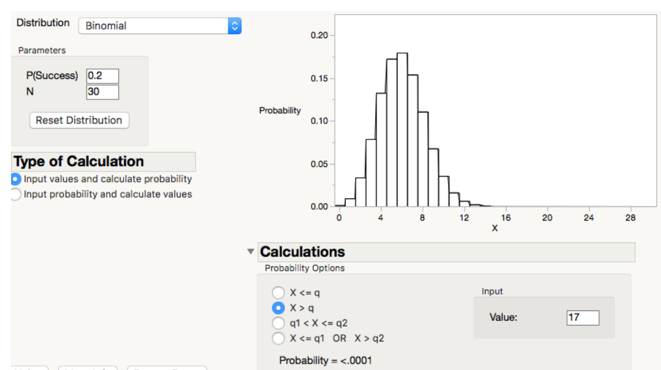
- The smaller this probability the less credibile his claim is. It should be stressed that this probability is **not the probability of his claim being true.**

Jack's Biology exam

- We calculate the chance of obtaining 18 or better out of 30, when only guessing.

- We note that the probability of scoring 18 or more out of the 30 in an exam is $P(S_{30} \geq 18|p = 0.2) = P(S_{30} = 18|p = 0.2) + \ldots + P(S_{30} = 30|p = 0.2) \approx 1.8 \times 10^{-6}$.





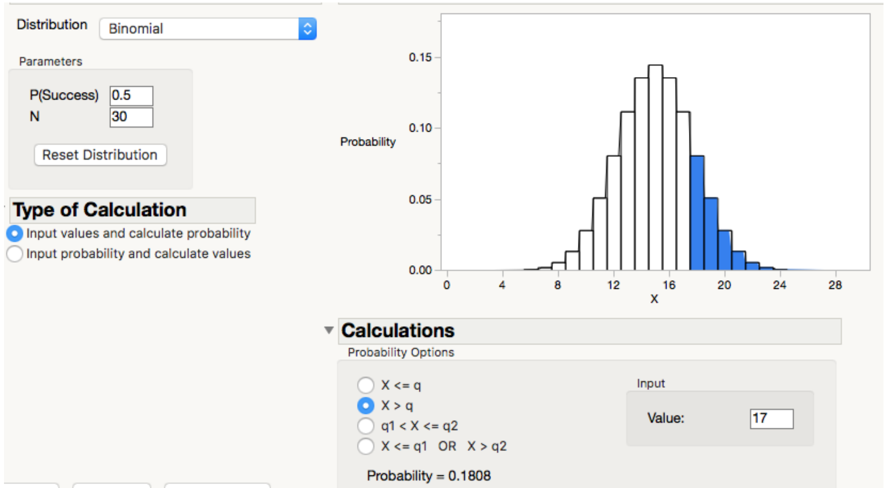
- This probability implies the chance of him guessing 18 or more is 0.0000018. Or in other words, if Jack were to do 10^7 exams (where he just guess all the answers), in about 18 of these exams he would score 18 or more points out of 30. This probability is called a **p-value**, it is the chance of observing the given data under the scenario that the null hypothesis is true.

  Rare events, such as this can happen. But a more plausible explanation for the score is that the alternative hypothesis, $p > 0.2$, is true. A score of 18 or more out of 30 is far more likely if the chance of answering a question correctly is greater than by random ($p > 0.2$).

- Conclusion; his score in his Biology exam **strongly suggests that he was not randomly guessing and the alternative hypothesis is true**.

The same data when $p = 0.5$

- The probability $p = 0.5$ means he is not randomly guessing but is making intelligent guesses based on some knowledge (but we assume independence between questions). The chance of scoring 18 or more out of 30 increases considerably (it is 18%). See the plot below.

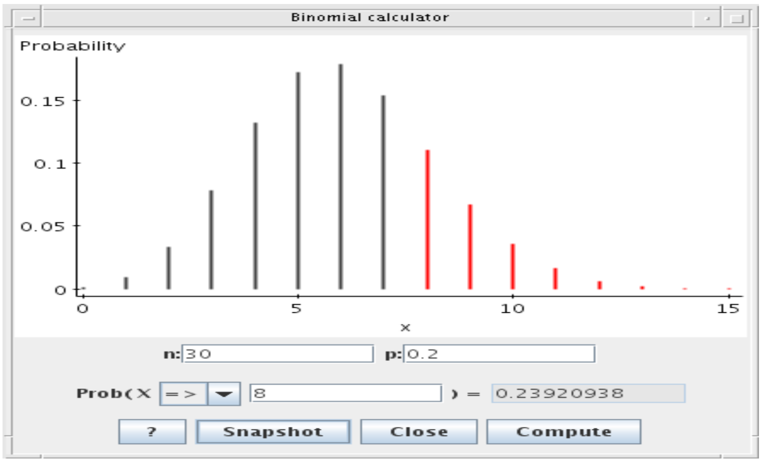


Since 0.00018% is extremely small and 18% is relatively large. The null seems unlikely and the alternative more plausible.

Jack's chemistry exam

- We test $H_0 : p \leq 0.2$ vs $H_A : p > 0.2$, based on his scoring 8 out of 30 in his chemistry exam.

  Using software we calculate $P(S_{30} \geq 8|p = 0.2) = 0.23$



- The probability of him scoring 8 or more by simply guessing is 0.23. In other words, if he did 100 exams in about 23 of them he would score at least 8 points out of 30.

The p-value for this test is 0.23 and it is not small. Therefore it is plausible he guessed. The score of 8 out of 30 is consistent with him guessing, therefore we cannot reject the null hypothesis.

- We cannot prove the null is true, as it is impossible to know whether he knew the answers to the 8 questions he answered correctly.

- Conclusion; there is no evidence in the data to reject the null.

- Even if the p-value were 100% we cannot accept the null. It simply states that the probability of the data being generated if the null were true is very high. However, the probability under a certain alternative could also be high. Thus based on the data we cannot make a decision about our hypothesis.

- A power analysis (which we do in a later lecture), will help us understand the implications of not rejecting the null (and what can be learnt about the alternative).

- Extremely important The p-value **does not** give the probability of the null being true. 1 - p-value **does not** give the probability of the alternative. This is a common misconception about p-values.

  Even with a p-value of 100% **we cannot** say the null is true!

Calculation practice

Let X_i be the probability the ith randomly selected person wins a game. $X_i = 0$ person losses $X_i = 1$ person wins.

$$P(X_i = 0) = 0.9 \qquad P(X_i = 1) = 0.1.$$

Let $S_4 = X_1 + X_2 + X_3 + X_4$.

(i) Calculate the probability two people out of four will win the game ($P(S_4 = 2)$).

(ii) Calculate the probability that two or less people will win the game ($P(S_4 \leq 2)$).

We construct all the possible different outcomes that can occur which give $S_4 = 2$.

Outcome	Per. 1	Per. 2	Per. 3	Per. 4	Probability
1	1	1	0	0	P(A)$=0.9^2 \cdot 0.1^2$
2	1	0	1	0	P(B)$=0.9 \cdot 0.1 \cdot 0.9 \cdot 0.1$
3	1	0	0	1	P(C)$=0.9 \cdot 0.1^2 \cdot 0.9$
4	0	1	1	0	P(D)$=0.1 \cdot 0.9^2 \cdot 0.1$
5	0	1	0	1	P(E)$=0.1 \cdot 0.9 \cdot 0.1 \cdot 0.9$
6	0	0	1	1	P(F)$=0.1^2 \cdot 0.9^2$
					$6 \cdot 0.1^2 \cdot 0.9^2$

- Remember each outcome is mutually exclusive to all the other outcomes, so $P(S_4) =$ P(A or B or C or D or E or F) $= P(A) + P(B) + P(C) + P(D) + P(E) + P(F)$.

- Since X_1, X_2, X_3, X_4 are all independent events. Then P(A)=$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0) = P(X_1 = 1)P(X_2 = 1)P(X_3 = 0)P(X_4 = 0) = 0.9^2 \cdot 0.1^2$.

- This gives $P(S_4 = 2) = 6 \cdot 0.9^2 \cdot 0.1^2$.

- Using the same argument we can show that $P(S_4 = 1) = 4 \cdot 0.9^3 \cdot 0.1$ and $P(S_4 = 0) = 0.9^4$.

- Therefore the probability that two or less win the game is the probability noone wins or one wins or two win:

$$
\begin{aligned}
P(S_4 \leq 2) &= P(S_4 = 0) + P(S_4 = 1) + P(S_4 = 2) \\
&= 0.9^4 + 4 \cdot 0.9^3 \cdot 0.1 + 6 \cdot 0.9^2 \cdot 0.1^2.
\end{aligned}
$$

Assumptions of a Binomial Experiment

The Binomial distribution is extremely useful. To use the binomial distribution the random sample (experiment) must satisfy the following assumptions:

(i) Each experiment (known as a Bernoulli trial) results in two outcomes (often refered as a success (yes) and failure (no)).

(ii) The probability of a success in each trial is equal to p.

(iii) The trials are independent.

See page 145 of Ott and Longnecker.

The binomial distribution: Example 4

The city wants to estimate the proportion of the population which are unemployed. A random sample of 5 people (without replacement) is taken from all the adults in a city. Each person is asked whether they are employed or not. We assume that the proportion of people unemployed is 0.1.

- Does our sample (experiment) satisfy the assumptions of a binomial distribution?

Solution

We recall that we observe X_1, X_2, X_3, X_4, X_5, where X_i be the answer of the ith person. $X_1 = 1$ if the person is unemployed and $X_1 = 0$ if employed. We want to check whether we have a Bernoulli experiment.

- Each experiment (person interviewed - bernoulli trial) results in a yes or no. So there are two outcomes.

- In this case the true p is

$$p = \frac{\text{Number of people in city who are unemployed}}{\text{Number of adults in city}},$$

and we suppose that $p = 0.1$. Clearly $P(X_i = 1) = 0.1$ and $P(X_i = 0) = 0.9$. Hence the probability of each draw is the same.

- The independence assumption is a little bit tricky. $P(X_2 = 1|X_1 = x)$ will not be exactly $P(X_2 = 1) = p$. The reason is that we have *removed* the observation X_1 from the population (since we sampled without replacement). So

$$P(X_2 = 1|X_1 = 1) = \frac{\text{Number of people in city who are unemployed} - 1}{\text{Number of adults in city} - 1},$$

Similarly

$$P(X_2 = 1|X_1 = 0) = \frac{\text{Number of people in city who are unemployed}}{\text{Number of adults in city} - 1},$$

Comparing $P(X_2 = 1|X_1 = 1)$ with $P(X_2 = 1)$ we see that they are not exactly the same. Recall for independence they need that $P(X_2 = 1|X_1 = 1) = P(X_2 = 1)$. However, if the population is large,

$P(X_2 = 1|X_1 = 1)$ and $P(X_2 = 1)$ are very close. In which case the independence assumption is close to holding. See Ott and Longnecker, Example 4.6 (page 145) for more details.

- In other words, when we do not replace the first observations. Knowledge of the first observations slightly changes the chance of the second observations. There is dependence. Though this dependence is very "small" is the sample is very small as compared with the population.

Example

- Consider a population of 1000 individuals. The random variable here is whether a randomly selected person is employed or not. Suppose that 250 people in the town are employed. Let X_1 be the employment status of the first person drawn and X_2 be employment status of second person drawn (without replacement). Then we see that

$$P(X_1 = \text{employed}) = \frac{250}{1000}, \quad P(X_2 = \text{employed}) = \frac{250}{1000}$$

and

$$P(X_2 = \text{employed}|X_1 = \text{employed}) = \frac{249}{999},$$

- We observe that $P(X_2 = \text{employed}) \neq P(X_2 = \text{employed}|X_1 = \text{employed})$, hence X_1 and X_2 are not independent. But because $250/100$ and $249/999$ are very close, they are "close to independent".

- Do not worry if you do not catch this argument. The main thing is if the sample size is small as compared with the population size then we have something close to independent samples and a Binomial experiment.

Additional facts: the mean and variance of a binomial

Recall that the number of successes out of n, denoted by S_n is a random variable taking values in $\{0, 1, \dots, n\}$ (eg. S_4 is the number of successes out of 4 and has the outcomes $\{0, 1, 2, 3, 4\}$). S_n has all the properties of a random variable, we can associate a probability to each outcome (the binomial distribution) and it has a probability plot. Since it has a probability plot, it must have a center and a spread, therefore it has a mean and a variance.

- The mean of a binomial is $n \times p$. This is very clear, for example if the chance of my getting a question correct is 80% and I answer 30 questions, on average I will get $0.8 \times 30 = 24$ question correct.

- The standard deviation of a binomial is $\sqrt{n \times p \times (1 - p)}$.