STATISTICS

A SHORT AND PAINLESS INTRODUCTION

Nils M Holm

Contents

Preface

Statistics is a fascinating subject but, unfortunately, there is little literature that teaches it in a way that is a accessible to mathematical laymen while still being precise and straight to the point.

On the one side, there are math textbooks that provide the interested reader with rigorous proof of every little detail, which may not be necessary for all students of humanities or natural sciences. On the other side, there are lengthy works that use entertaining language and cute comic figures to try to get their point across.

This book has neither rigorous proofs nor cute characters. It strives to build a solid foundation for people who use statistics as a tool. After finishing this book, the reader should be able to digest more complete works in the field of statistics without too much difficulty.

Topics covered in this brief volume include basic probability, probability functions and distributions, confidence intervals, linear regression, correlation, and hypothesis testing.

The final chapters describe numerical methods for computing statistical functions. They are optional for most students, but may be of interest to the mathematically inclined reader.

There are no exercises, but at some points questions are asked and at those points the reader is invited to put aside the book and try to find their own solution before reading on.

The matter of the book progresses rather quickly, so the reader is advised not to skip sentences or paragraphs. Doing so would probably (!) complicate the comprehension of later parts of the text.

Enjoy the tour through the foundations of statistics!

Nils M Holm, July 2016

Probability

Basics

Probability is an *estimate* predicting how often an *event* will occur given a fixed number of *trials*. For example, when a coin is tossed 10 times, “heads” will probably show up *about* 5 times. So the probability of “heads” is 50% or

$p = 0.5$

In statistics, probabilities are expressed as a real number $p \in \mathbf{R}$ where $0 \leq p \leq 1$, i.e. p is in the interval $[0, 1]$. *Impossibility* (an event will never occur) is denoted by $p = 0$ and *certainty* (an event will always occur) is represented by $p = 1$.

Note that getting 7 “heads” and 3 “tails” when tossing a coin 10 times does not violate the prediction of getting “heads” half of the time! The probability of $p = 0.5$ for getting “heads” is only the *most probable outcome* of tossing a coin repeatedly.

As the number of trials (coin tosses) increases, the actual distribution of heads and tails will converge towards the expectation of $p = 0.5$ for “heads”.

When tossing two coins at the same time, there are four possible *outcomes* (H indicates “heads” and T indicates “tails”):

$HH \quad HT \quad TH \quad TT$

Each outcome has the same probability of $p = 0.25$. This can be expressed using the *probability function* P as follows:

$P(HH) = 0.25$
$P(HT) = 0.25$
$P(TH) = 0.25$
$P(TT) = 0.25$

Meaning: the probability of two times “heads” is $p = 0.25$, etc.

The set of all possible outcomes is called the *sample space* S. In the above example, this would be

$$S = \{HH, HT, TH, TT\}$$

The probability of S is $P(S) = 1$, because one of the above events *must* occur in every trial (we are assuming an "ideal" coin that cannot get stuck on its edge or disappear in a storm drain).

So if $S = A_1 \cup \cdots \cup A_n$, then $P(A_1) + \cdots + P(A_n) = P(S) = 1$, where $A \cup B$ denotes the *union*, or logical "or", of two events, i.e. either A or B or both A and B happens.

The notation A' (sometimes also $\bar{A}$) is called the *complement* of A. It indicates that an event A does *not* occur. The probability $P(A')$ is $1 - P(A)$ for any A. For example, the probability of not getting two times heads when tossing two coins would be

$$P(\overline{HH}) \ = \ 1 - P(HH) \ = \ 1 - 0.25 \ = \ 0.75$$

and the probability of none of the events in the sample space happening would be $P(S') = 1 - 1 = 0$.

When drawing cards from a standard deck of 32 cards, the probability of drawing an "ace" would be $P(A) = \frac{4}{32} = \frac{1}{8}$, because there are 4 aces in the standard deck. The probability of drawing a red card from the deck would be $P(B) = \frac{16}{32} = \frac{1}{2}$, because there are 16 red and 16 black cards in the deck. The sample space in this case would be the entire deck.

The probability of drawing a "red ace" would be

$$P(A \cap B) \ = \ P(A) \cdot P(B) \ = \ \frac{1}{8} \cdot \frac{1}{2} \ = \ \frac{1}{16} \ = \ 0.0625$$

Meaning: the probability of A *and* B happening *at the same time* (i.e. in the same trial) is 0.0625.

The probability of the *intersection*, or logical "and", $P(A \cap B)$ of two *independent* events A and B is calculated by multiplying their probabilities. More on this later (pg 14).

The probability of drawing an "ace" *or* a red card is:

$$P(A \cup B) \ = \ P(A) + P(B) - P(A \cap B) \ = \ \frac{1}{8} + \frac{1}{2} - \frac{1}{16} \ = \ 0.5625$$

Meaning: the probability of *either* A *or* B *or* $A \cap B$ happening in the same trial is 0.5625.

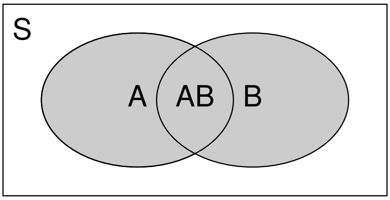
When calculating the probability of the *union* $P(A \cup B)$ of two events A and B, the probability of the intersection $P(A \cap B)$ has to be subtracted, because otherwise it would be duplicated (if it is non-zero).

Of course, if the intersection of A and B is empty, its probability does not have to be subtracted, so $P(A \cup B) = P(A) + P(B)$, *iff* A and B are mutually exclusive, i.e. iff the events A and B *cannot* occur in the same trial.

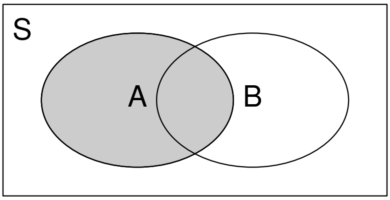
(Note: “iff” is a common abbreviation for the bidirectional “if”, i.e. “if and only if”.)

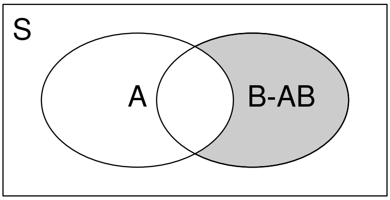
In the above example, $P(A)$ (aces) includes two red cards and $P(B)$ (red cards) includes two aces, giving an “overlap” of 4 cards. However, there are only 2 red aces in the deck, so half of the overlap has to be eliminated by subtracting $P(A \cap B)$.

This is probably best demonstrated using a Venn diagram (each ellipse denotes an event and the overlap of ellipses denotes the intersection of events):



Both the event A and B would contribute to the intersection AB, thereby duplicating it, so one of the sets (ellipses) has to lose its intersection part before adding it:





Summary

$0 \le p \le 1$ for any probability p.

$0 \le P(A) \le 1$ for any event A.

$P(A') = 1 - P(A)$ for any event A.

$P(S) = 1$ for any sample space S.

If $S = \bigcup_{i=1}^{n} A_i$, then $P(S) = \sum_{i=1}^{n} P(A_i) = 1$.

$$P(A \cap B) = P(A) \cdot P(B), \text{ iff } A \text{ and } B \text{ are independent.}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

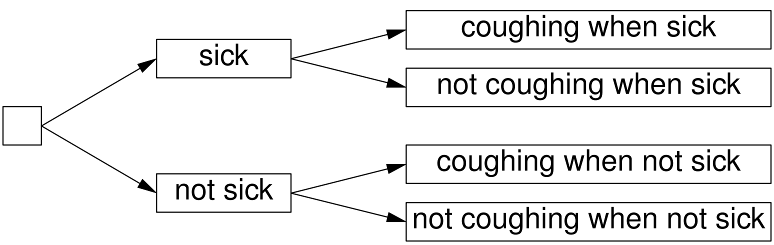
$$P(A \cup B) = P(A) + P(B), \text{ iff } A \text{ and } B \text{ are mutually exclusive.}$$

Conditional Probability

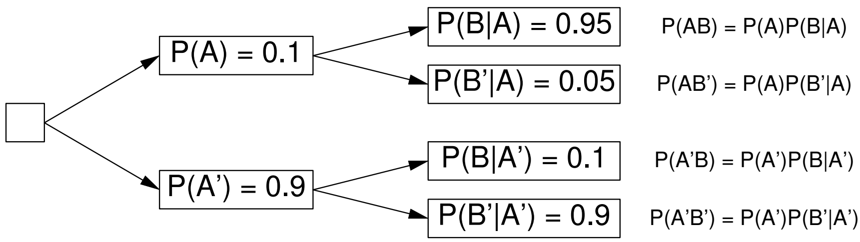
Imagine it is flu season and

- the probability of a random person having a cold is $p = 0.1$
- the probability of a person coughing while having a cold is $p = 0.95$ (5% may not cough and still be sick)
- the probability of a person not coughing while not having a cold is $p = 0.9$ (10% might cough for different reasons)

The following tree diagram can be constructed from this scenario:



The notation $P(B|A)$ denotes the *conditional probability* of "B *given* A", i.e. the probability of B in the case where it is already known that A has happened. Using this notation, the diagram can be populated with probabilities:



For example, $P(A)$ denotes the probability of a person being sick, $P(B|A)$ denotes the probability of a person coughing *given* they are sick, and $P(AB) = P(A \cap B) = P(A) \cdot P(B|A)$ is the probability of a person coughing *and* being sick $(0.1 \cdot 0.95 = 0.095)$.

Note that the probability $P(A \cap B)$ is given as $P(A) \cdot P(B|A)$ here, while it was given as $P(A) \cdot P(B)$ earlier in this text (pg 11). In the example given here, A and B are *dependent*, because $P(B) \neq P(B|A)$.

In fact, two events A and B are *independent* if, and only if, $P(B) = P(B|A)$. That is, $P(B)$ is the same, no matter whether A has is given or not.

The probability $P(B)$ of a person coughing, although not explicitly stated in the data, can be inferred from the diagram. It is the combined probability of a person coughing, no matter if they have a cold or not:

$$\begin{aligned} P(B) &= P(B \cap A) \ + \ P(B \cap A') \\ &= P(A) \cdot P(B|A) \ + \ P(A') \cdot P(B|A') \\ &= 0.1 \cdot 0.95 + 0.9 \cdot 0.1 \\ &= 0.185 \end{aligned}$$

Note that $P(B \cap A)$ and $P(B \cap A')$ are mutually exclusive, because A and A' cannot happen in the same trial. (E.g. a person cannot be sick and not sick at the same time.)

Reverse Conditional Probability

A more interesting question in the flu season might be: “Given that someone is coughing, what is the probability that they have a cold?” I.e.: what is $P(A|B)$?

The probability of a person coughing *because* they have a cold equals the *proportion* of people who cough while having a cold and those who cough at all (for whatever reason):

$$P(A|B) \;=\; \frac{\text{coughing and sick}}{\text{coughing}} \;=\; \frac{P(A\cap B)}{P(B)}$$

Inserting the formulae for $P(A\cap B)$ and $P(B)$ from above gives:

$$\frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A)+P(A')\cdot P(B|A')}$$

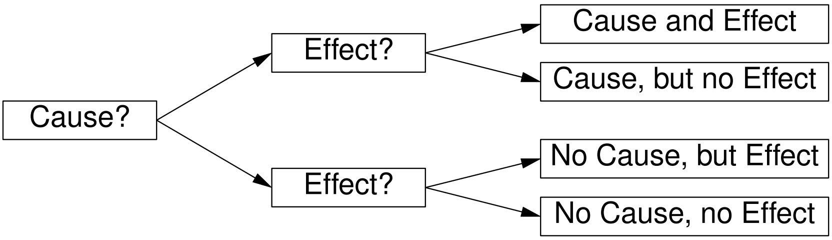
All probabilities that appear in this formula can be extracted from the tree diagram. The above formula is widely known as *Bayes' Theorem* or *Bayes' Rule*.

Substituting values for probability functions finally gives:

$$\frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A)+P(A')\cdot P(B|A')} \;=\; \frac{0.1\cdot 0.95}{0.1\cdot 0.95+0.9\cdot 0.1} \;=\; \frac{19}{37}$$

The probability of someone coughing because they have a cold in said flu season is just $p=0.514$ — so results from the above test would only be slightly more significant than tossing a coin.

The basic idea behind *reverse conditional probability* is as follows:



We may observe an *effect*, and that effect may or may not have a specific *cause*. Given the probability of "effect given cause" ($P(B|A)$) and the probability of "no effect given no cause" ($P(B'|A')$) as well as the probability of the cause in general ($P(A)$), what is the probability of "cause given effect"? I.e. what is the probability that the observation of the effect was triggered by the specific cause? This question is answered by Bayes' Theorem.

When using reverse conditional probability (RCP) to evaluate test results, $P(B|A)$ is called the *sensitivity* of the test and $P(B'|A')$ is called its *specificity*. A test is *sensitive*, if it catches a lot of positives (i.e. has few false negatives). It is *specific*, if it catches few negatives (i.e. has few false positives).

A *false positive* occurs when a test wrongly delivers a positive result. Analogously, a *false negative* is a wrong negative result.

The above “cold test” has a sensitivity of $p = 0.95$ and a specificity of $p = 0.9$. This sounds good, but the reliability of the test is low, because a positive only indicates a $p = 0.514$ chance for the specific cause.

RCP depends a lot on the *prior probability* (or just “prior”), i.e. the probability of the cause in general, $P(A)$. When $P(A) = 1$, then

$$\frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A)+P(A')\cdot P(B|A')}=\frac{1\cdot P(B|A)}{1\cdot P(B|A)+0\cdot P(B|A')}=\frac{P(B|A)}{P(B|A)}$$

So the test result is always $p = 1$. Similarly, when the prior is $P(A) = 0$, the test will always be negative.

Above cold test would not be as bad, if the *prevalence* (the medical term for the prior) was higher, i.e. if more people had a cold in the first place. Given $P(A) = 0.5$:

$$\frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A)+P(A')\cdot P(B|A')}=\frac{0.5\cdot 0.95}{0.5\cdot 0.95+0.5\cdot 0.1}=\frac{19}{21}$$

So given a prevalence of 50%, a coughing person would indicate a cold in about 90% of the observed cases.

The smaller the prior is, the greater the sensitivity and the specificity of a test have to be in order for the test result to be significant.

Summary

$P(A\cap B) = P(A)\cdot P(B|A)$ if A and B are dependent.

$P(B) = P(B|A)$ iff A and B are independent.

$P(B) = P(A)\cdot P(B|A) + P(A')\cdot P(B|A')$

Probability Distributions

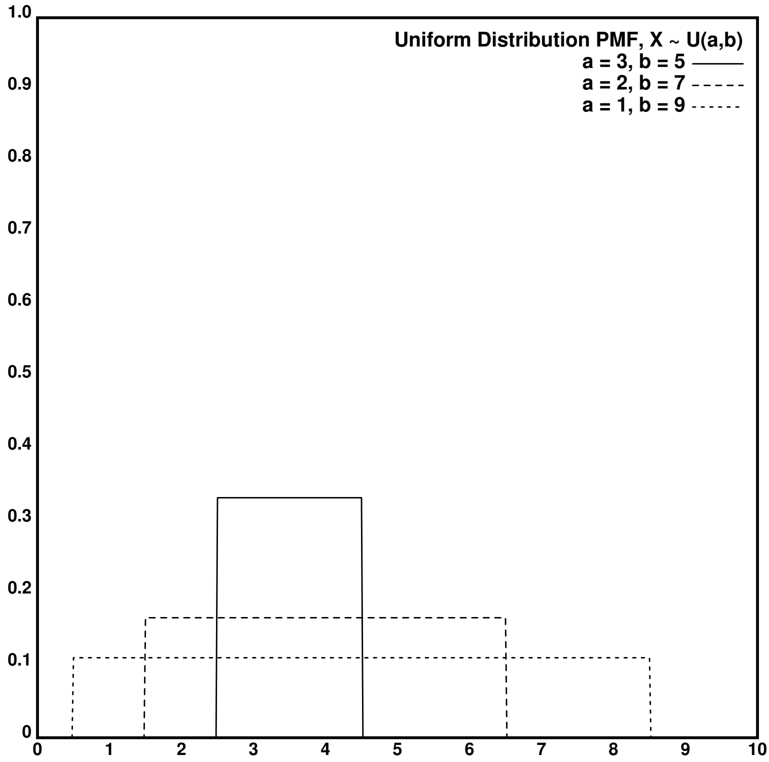
Uniform Distribution

$X \sim U(a,b)$	
PMF	$\begin{cases} 0 & \text{if } x < a \\ \dfrac{1}{b-a+1} & \text{if } a \le x \le b \\ 0 & \text{if } x > b \end{cases}$
CDF	$\begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a+1}{b-a+1} & \text{if } a \le x \le b \\ 1 & \text{if } x > b \end{cases}$
Parameters	$a, b \in \mathbf{Z}, a \le b$: range
μ	$\dfrac{a+b}{2}$
σ^2	$\dfrac{(b-a+1)^2-1}{12}$
Skewness (γ_1)	0

Question answered

PMF: what is the probability of an event x happening, given a constant probability?

CDF: what is the probability of any event in the range from a to x happening?



Examples

The probability of getting a specific face when rolling a six-sided die follows the uniform distribution $X \sim U(1,6)$, so the probability of getting a "3" is

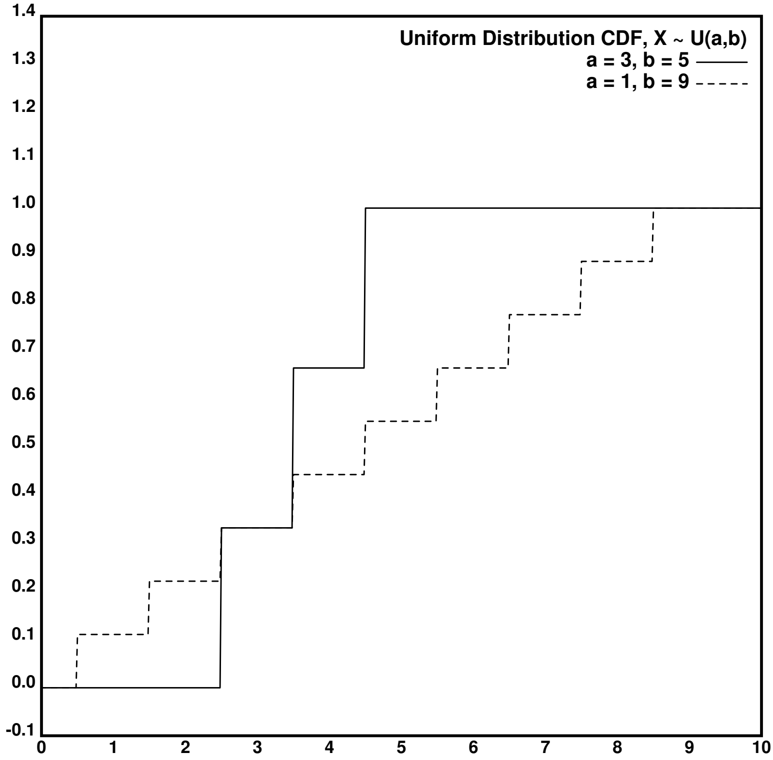
$$P(X = 3) = \frac{1}{b-a+1} = \frac{1}{6-1+1} = \frac{1}{6}$$

The probability of getting a "1", "2", or "3" is:

$$P(X \le 3) = \frac{x-a+1}{b-a+1} = \frac{3-1+1}{6-1+1} = \frac{1}{2}$$

The probability of getting at least a "3" is:

$$P(X \ge 3) = 1 - P(X \le 2) = 1 - \frac{x-a+1}{b-a+1} = 1 - \frac{2-1+1}{6-1+1} = \frac{2}{3}$$


Uniform Distribution CDF, X ~ U(a,b)
a = 3, b = 5
a = 1, b = 9
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0
1
2
3
4
5
6
7
8
9
10

Geometric Distribution

$X \sim Geo(p)$	
PMF	$q^{x-1} \cdot p$
CDF	$1 - q^x$
Parameters	$p \in [0, 1]$: probability of success
	q: $1 - p$
	$x \in \mathbf{N}_0$: number of trials
μ	$\frac{1}{p}$
σ^2	$\frac{q}{p^2}$
Skewness (γ_1)	$\frac{2-p}{\sqrt{q}}$

Questions answered

PMF: given x independent trials, all with equal probability of success p, what is the probability of *exactly* one success after $x - 1$ failures?

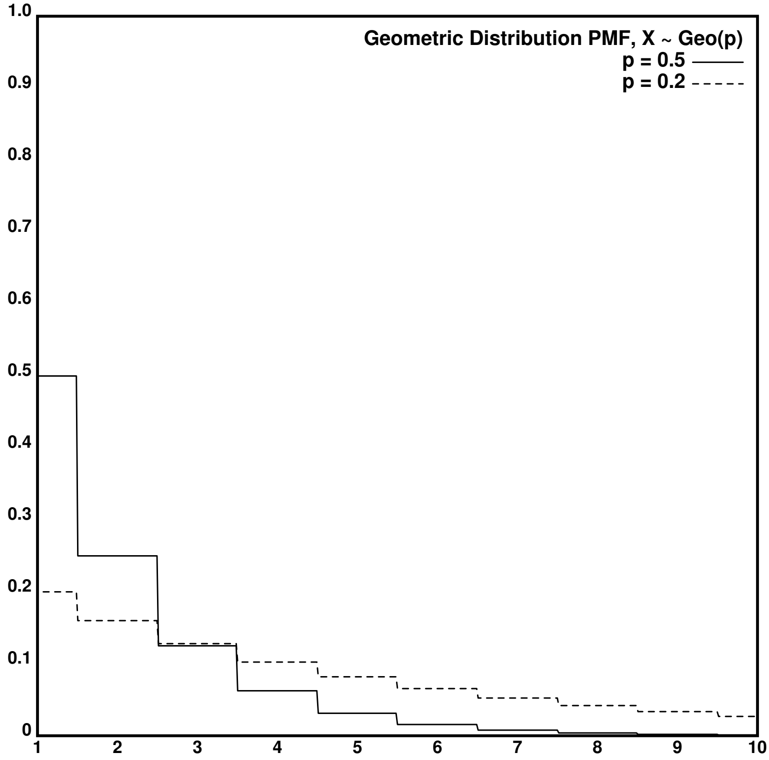
CDF: given x independent trials, what is the probability of *at least* one success?

Examples

The probability of getting the first "six" in the n^{th} subsequent roll of a six-sided die follows the geometric distribution $X \sim Geo(\frac{1}{6})$. The probability of getting a six ($p = \frac{1}{6}$) in the $x = 3^{rd}$ toss of a die is:

$$P(X = 3) = q^{x-1} \cdot p = \left(1 - \frac{1}{6}\right)^2 \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^2 \cdot \frac{1}{6}$$

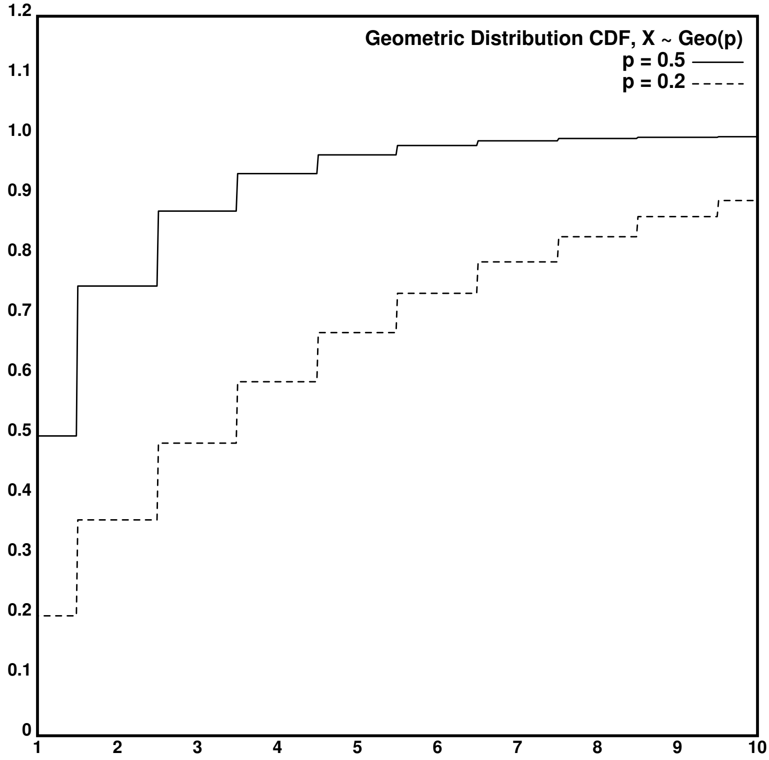
$$= \frac{25}{36} \cdot \frac{1}{6} = \frac{25}{216} \approx 0.116$$



The probability of getting at least one six in three tosses is:

$$P(X \le 3) = 1 - q^x = 1 - \left(1 - \frac{1}{6}\right)^3 = 1 - \left(\frac{5}{6}\right)^3$$

$$= 1 - \frac{125}{216} = \frac{91}{216} \approx 0.421$$


Geometric Distribution CDF, X ~ Geo(p)
p = 0.5
p = 0.2
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10

Binomial Distribution

$X \sim B(n, p)$	
PMF	$\binom{n}{x} \cdot p^x \cdot q^{n-x}$
CDF	$\sum_{i=0}^{x} \binom{n}{i} \cdot p^i \cdot q^{n-i}$
	$I_q(n-x, 1+x)$
Parameters	$n \in \mathbf{N}_0$: number of trials
	$p \in [0, 1]$: probability of success
	q: $1-p$ (probability of failure)
	$x \in \mathbf{N}_0, x \le n$: number of successes
μ	np
σ^2	npq
Skewness (γ_1)	$\frac{q-p}{\sqrt{npq}}$
Approximations	$N(np, npq)$ for $np > 5, nq > 5$
	$Poi(np)$ for $n \ge 50, p < 0.1$

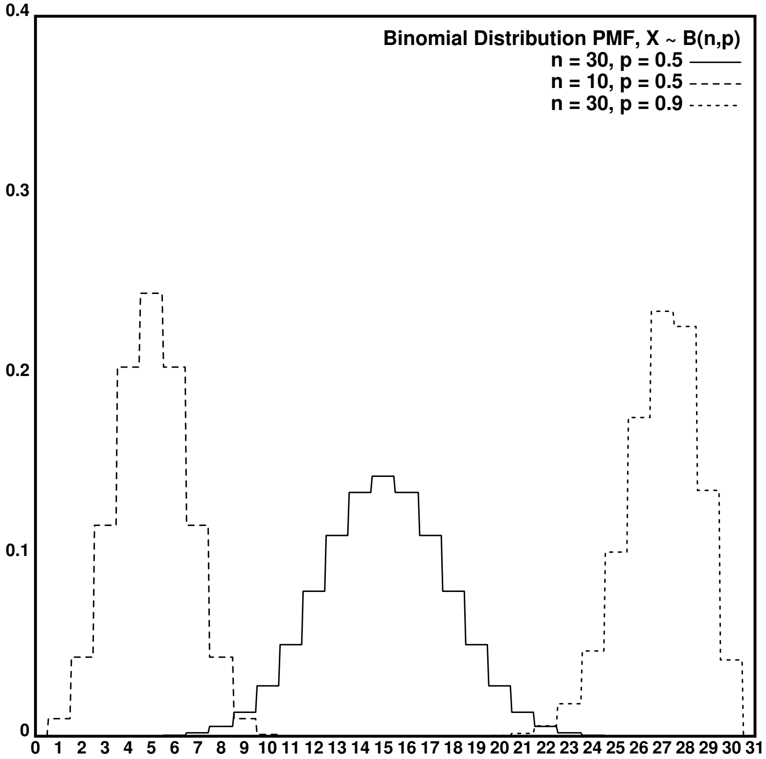
Questions answered

PMF: what is the probability of *exactly* x successes in n independent trials with a success probability of p?

CDF: what is the probability of *up to* x successes in n independent trials?

Examples

The probability for x out of 5 children being girls (or boys) follows the binomial distribution $X \sim B(5, 0.5)$ given equal chances for a child being a girl or boy ($p = 0.5$).



Given this distribution, the probability for a couple having exactly $x = 3$ girls is:

$$P(X=3) = \binom{n}{x} \cdot p^x \cdot q^{n-x} = \binom{5}{3} \cdot 0.5^3 \cdot (1-0.5)^{5-3}$$

$$= \binom{5}{3} \cdot 0.5^3 \cdot 0.5^2 = 10 \cdot 0.125 \cdot 0.25 = 0.3125$$

All other factors being equal, the probability of *up to* three of the children being girls is:

$$P(X \leq 3) = \sum_{i=0}^{x} \binom{n}{i} \cdot p^i \cdot q^{n-i} = \sum_{i=0}^{3} \binom{5}{i} \cdot 0.5^i \cdot 0.5^{n-i}$$

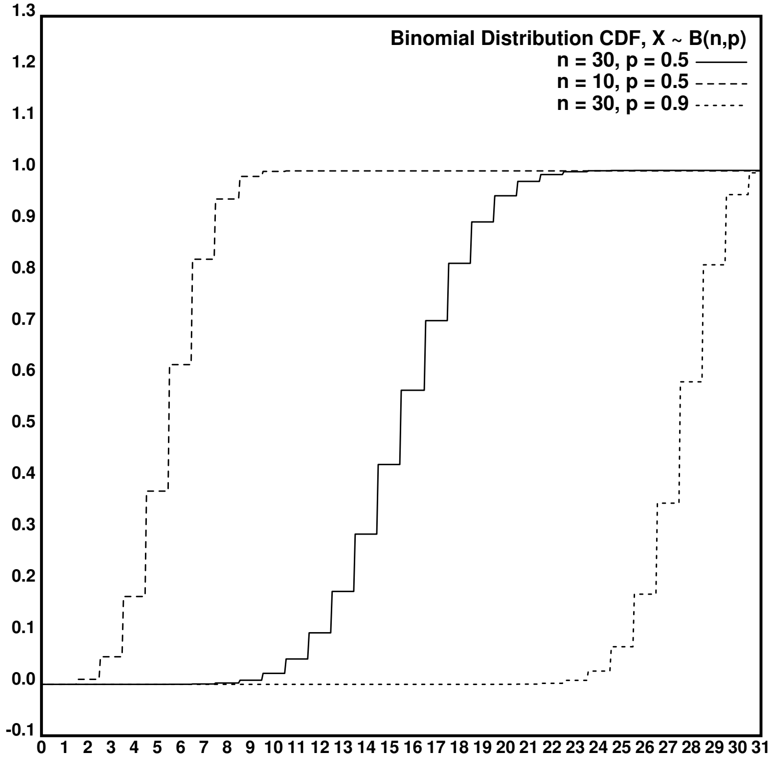
$$= \binom{5}{0} \cdot 0.5^0 \cdot 0.5^5 + \binom{5}{1} \cdot 0.5^1 \cdot 0.5^4 + \binom{5}{2} \cdot 0.5^2 \cdot 0.5^3 + \binom{5}{3} \cdot 0.5^3 \cdot 0.5^2$$

$$= 0.03125 + 0.15625 + 0.3125 + 0.3125 = 0.8125$$

A more effective way to compute the above would be:

$P(X \leq 3) = I_q(n-x, 1+x) = I_{0.5}(2, 4) = 0.8125$

where $I_q(a, b)$ denotes the regularized incomplete B (beta) function (see page 108).


Binomial Distribution CDF, X ~ B(n,p)
n = 30, p = 0.5
n = 10, p = 0.5
n = 30, p = 0.9
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Index