STA257 All Notes 2016

Neil Montgomery

admin

contact, websites, notes

date format	ISO8601
instructor	Neil Montgomery
email	neilm@mie.utoronto.ca
office	BA8137
office hours	any time
website	portal (announcements, grades, etc.)
github	https://github.com/sta257-fall-2016 (lecture material, code, etc.)

Official lecture notes are HTML.

PDFs will be uploaded before classes for people who like to annotate them during lecture on a PED. But they will never be updated after class.

evaluation, book, tutorials

what	when	how much
midterm 1	2016-10-03 (OOPS!) during class time	25%
midterm 2	2016-11-14 during class time	25%
exam	TBA	50%

The book is Mathematical Statistics and Data Analysis, 3rd Edition by Rice (also used in STA261).

I will suggest exercises from this book each week.

Your TA will work through some of them in tutorial each week.

other resources

Other (free, downloadable PDF through library) similar books:

Introduction to Probability with Statistical Applications by Schay (similar level) An Intermediate Course in Probability by Gut (more advanced) Introduction to probability models by S. Ross (a bit more engineer-y)

Avoid books with titles like Introduction to Probability and Statistics For <Some Certain Specific Type of Student or Career>

A very nice (free PDF through library) book for those really interested in the mathematics of all this is: Elementary Analysis by K. Ross. (Subtitled "The Theory of Calculus".)

Wikipedia, youtube, stackexchange, google. Lots of good stuff. Lots of bad stuff.

meanings of probability

no need to write this down

Physical, long term relative frequency, "repeated experiments", "frequentist", "propensity"

Evidential, subjective, "Bayesian", inductive

Visit Department of Philosophy for more information.

Whatever the interpretation, the mathematical rules are the same, based on axioms that define how probabilities can be assiged to events.

axiomatic approach to probability

elements and sets

We can think of an element ω belonging to a set A. We can think of sets A and B along with a universal set S. We have the following notions, and more:

Membership $\omega \in A$

Union "or" $A \cup B$; Intersection "and" $A \cap B$; works for infinitely many

Complement $A^c = \{w \in S : w \notin A\}$

Empty set has no elements: $\emptyset$

Disjointness: $A \cap B = \emptyset$ (notice: not a probability concept)

Subset: $A \subseteq B$ (and "proper subset")

Set difference: $A \backslash B = A \cap B^c$

sample space

Probability starts with a sample space S, a collection with all possible outcomes of the random process. Often cumbersome and arbitrary; mainly used this week. Examples:

Coin toss: $\{H, T\}$

Picking a card: $\{A\spadesuit, A\heartsuit, A\diamondsuit, A\spadesuit\}$

Toss two coins: $\{HH, HT, TH, TT\}$. Or possibly: $\{\{H, H\}, \{H, T\}, \{T, T\}\}$

A race among 8 horses: ?

Toss a coin until a head appears: $\{H, TH, TTH, TTTH, \ldots\}$

Pick a real number between 0 and 1 "uniformly": $(0, 1)$ (A "continuous" sample space.)

event

An event is a collection of outcomes; equivalently a subset of the sample space S.

Naming conventions: Roman letter near the beginning $A, B, C, \dots$ or $A_1, \dots, A_n$ or $A_1, A_2, A_3, \dots$ as required.

Many examples possible from the example sample spaces.

it's really more complicated than that

This is an elementary course, so we will not concern ourselves with the fact that not all subsets of a sample space are allowed to be called "events".

Really an event has to be a "suitable" collection of outcomes.

For finite and countable (i.e. "list-able") sample spaces, in fact all events are "suitable".

But not for uncountable sample spaces, such as $(0, 1)$.

The "space" of suitable events can be called $\mathcal{A}$.

the probability axioms

A probability measure is a real-valued function (usually called) P. Its domain is $\mathcal{A}$, a space of suitable events in S. In addition, it has the following properties:

1. $P(S) = 1$
2. $P(A) \geq 0$ for all events $A \in \mathcal{A}$.
3. If $A_1, A_2, A_3, \ldots$ is a disjoint collection of events, then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The last property is called "$\sigma-$additivity".

There is a notion of "probability triple" $(S, \mathcal{A}, P)$ when needed to be absolutly clear (which isn't often in this course.)

the axiomatic approach

In the fussiest possible treatment, the first question is: are these axioms consistent, which is the same as asking "Are there any probability measures at all?"

Theorem 0: the axioms are consistent.

Proof: $\ldots$

Advanced note…when the sample space is something like $S = (0, 1)$ and if we were to allow $\mathcal{A}$ to be the collection of all subsets of S, then the axioms are inconsistent.

Added after class due to interest from many students…the advanced note should be understood to include also the desire for the "uniform probability" on S - in other words the probability to lie in (a, b) with $0 < a < b < 1$ to be $b - a$.

everyday properties of P

Continuing with total and absolute fussiness:

Theorem 1: $P(\emptyset) = 0$

Proof: $\ldots$

Theorem 2: If A_1 and A_2 are disjoint then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

Proof: $\ldots$ (Note added after class…from now on we don't need the "silly" infinite number of ∅ in any of the proofs.)

Theorem 2a: If $A_1, A_2, \ldots, A_n$ are disjoint then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ (called "finite additivity")

Proof: "Induction" (Note: the book lists this Theorem as an "axiom", which is not technically wrong but…)

more everyday properties of P, with proofs

Theorem 3: $P(A^c) = 1 - P(A)$

Theorem 4: If $A \subseteq B$ then $P(A) \leq P(B)$. "P is monotone."

Theorem 4a: $P(B \backslash A) = P(B) - P(B \cap A)$ (Proof left as exercise)

Theorem 4b: If $A \subseteq B$ then $P(B \backslash A) = P(B) - P(A)$. (Proof left as exercise. Note that any proof left as an exercise is always allowed to use theorems already proven, and anything from general maths and calculus.)

Theorem 5: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (and generalizations)

This is admittedly getting dull.

a non-everyday application of this axiomatic approach to P

towards showing the "continuity" of P

The culmination of our axiomatic approach will be to define the notion of "continuity" for P and prove that the defined property holds.

Recall from the prerequiste the notion of a continuous function. There are several equivalent definitions, one of which uses left- and right-continuity.

A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is left-continuous at x if for any non-decreasing sequence $x_1 \leq x_2 \leq x_3 \leq \ldots$ that converges to x, then $f(x_i) \rightarrow f(x)$.

(Right continuity is the same but with a non-increasing sequence.)

f is continuous at x if it is left- and right-continuous at x.

f is continuous if it is continuous at every point in its domain.

increasing sequences of events that converge

The domain of P is a collection of events $\mathcal{A}$. We need a notion of the following for events:

$$A_1, A_2, \dots \text{ increases to } A$$

Definition: $A_n \nearrow A$ means $A_i \subseteq A_{i+1}$ and $\bigcup_{i=1}^{\infty} A_i = A$

Example: Consider $S = (0, 1)$. Let $A_n = \left(0, \frac{1}{2} - \frac{1}{2^{n+1}}\right)$ for $n \geq 1$ and $A = \left(0, \frac{1}{2}\right)$

What about the probabilities of these events under the uniform model?

the continuity theorem

Theorem 6 (The Continuity Theorem): If A_n and A are events and $A_n \nearrow A$, then $P(A_n) \longrightarrow P(A)$.

Proof: $\dots$

This is analogous to left-continuity. There is also a right-continuity:

Corallary: Suppose A_n and A are events in $\mathcal{A}$ with $A_i \supseteq A_{i+1}$ and $\bigcap_{i=1}^{\infty} A_i = A$. Then $P(A_n) \longrightarrow P(A)$.

Proof: The Continuity Theorem, a de Morgan's Law, and "Theorem 3".

Something to try if you like: finite additivity together with The Continuity Theorem implies $\sigma-$additivity.

application to the continuous sample space example

Reconsider the uniform pick on $S = (0, 1)$, where the probability of choosing a number in any $0 < a < b < 1$ is $b - a$.

What is the probability of choosing exactly $\frac{1}{2}$?

Let A be the event that the number chosen is rational. What is P(A)?

some computations for finite and countable sample spaces

finite and countable S in general

Starting with:

$$
\begin{aligned}
S &= \{\omega_1, \ldots, \omega_n\} \qquad \text{(finite), or,} \\
S &= \{\omega_1, \omega_2, \omega_3, \ldots\} \qquad \text{(countable)}
\end{aligned}
$$

then a valid probability can always be based on $P(\{\omega_i\}) = p_i$ with $0 \leq p_i \leq 1$ and $\sum p_i = 1$.

An important special case for finite S is the uniform case: $p_i = \frac{1}{n}$.

In this case many problems can be solved by counting the number of outcomes in S and counting the number of outcomes in an event.

Some people enjoy these problems. Others don't. Fortunately for you, I do not!

permutations and combinations

At the very least we'll need to recall (or learn!) these.

Number of ways to choose k items out of n where order matters:

$$_nP_k = \begin{cases} 0 & \text{if k > n,} \\ \frac{n!}{(n-k)!} & \text{otherwise.} \end{cases}$$

and when order doesn't matter:

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

Two classic examples: "The Birthday Problem" and "Lotto"

conditional probability

partial information

I'll role a six-sided die. $S = \{1, 2, 3, 4, 5, 6\}$. Consider these events:

$$\begin{aligned} A &= \{2, 5\}, \\ B &= \{2, 4, 6\}, \\ C &= \{1, 2\}. \end{aligned}$$

So $P(A) = \frac{2}{6} = \frac{1}{3}$.

What if I peek and tell you "Actually, B occurred". What is the probabality of A given this partial information? It is $\frac{1}{3}$.

I roll the die again, peek, and tell you "Actually, C occurred". Now the probability of A is $\frac{1}{2}$.

Intuitively we used a "sample space restriction" approach.

elementary definition of conditional probability

Given B with $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

"The conditional probability of A given B"

The answers for the previous example coincide with the intuitive approach.

Theorem 7: For a fixed B with $P(B) > 0$, the function $P_B(A) = P(A|B)$ is a probability measure.

Proof: exercise.

useful expressions for calculation - I

$P(A \cap B) = P(A|B)P(B)$ often comes in handy.

Consider the testing for, and prevalence of, a viral infection such as HIV.

Denote by A the event "tests positive for HIV", and by B the event "is HIV positive."

For the ELISA screening test, $P(A|B)$ is about 0.995. The prevalence of HIV in Canada is about $P(B) = 0.00212$.

useful expressions for calculation - II

Take some event B. The sample space can be divided in two into B and B^C.

This is an example of a partition of S, which is generally a collection $B_1, B_2, \dots$ of disjoint events (could be infinite) such that $\bigcup_i B_i = S$.

Theorem 8: If $B_1, B_2, \dots$ is a partition of S with all $P(B_i) > 0$, then

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

Proof: …

Continuing with the HIV example, suppose we also know $P(A|B^c) = 0.005$ ("false positive").

We can now calculate $P(A)$.

useful expressions for calculation - III

Much to my amusement, Theorem 8 gets a grandiose title: "THE! LAW! OF! TOTAL! PROBABILITY!!!"

Now, in the HIV example, we also might be interested in $P(B|A)$, the chance of an HIV+ person testing positive.

A little algebra:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

In our example this is $\frac{0.0021094}{0.0070988} = 0.2971.$

Bayes' rule

Theorem 9: If $B_1, B_2, \dots$ is a partition of S with all $P(B_i) > 0$, then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)}$$

permutations and combinations

At the very least we'll need to recall (or learn!) these.

Number of ways to choose k items out of n where order matters:

$${}_nP_k = \begin{cases} 0 & \text{if k > n,} \\ \frac{n!}{(n-k)!} & \text{otherwise.} \end{cases}$$

and when order doesn't matter:

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

Two classic examples: "The Birthday Problem" and "Lotto"

conditional probability

partial information

I'll roll a six-sided die. $S = \{1, 2, 3, 4, 5, 6\}$. Consider these events:

$$\begin{aligned} A &= \{2, 5\}, \\ B &= \{2, 4, 6\}, \\ C &= \{1, 2\}. \end{aligned}$$

So $P(A) = \frac{2}{6} = \frac{1}{3}$.

What if I peek and tell you "Actually, B occurred". What is the probabality of A given this partial information? It is $\frac{1}{3}$.

I roll the die again, peek, and tell you "Actually, C occurred". Now the probability of A is $\frac{1}{2}$.

Intuitively we used a "sample space restriction" approach.

elementary definition of conditional probability

Given B with $P(B) > 0$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

"The conditional probability of A given B"

The answers for the previous example coincide with the intuitive approach.

Theorem 7: For a fixed B with $P(B) > 0$, the function $P_B(A) = P(A \mid B)$ is a probability measure.

Proof: exercise.

useful expressions for calculation - I

$P(A \cap B) = P(A \mid B)P(B)$ often comes in handy.

Consider the testing for, and prevalence of, a viral infection such as HIV.

Denote by A the event "tests positive for HIV", and by B the event "is HIV positive."

For the ELISA screening test, $P(A \mid B)$ is about 0.995. The prevalence of HIV in Canada is about $P(B) = 0.00212$.

useful expressions for calculation - II

Take some event B. The sample space can be divided in two into B and B^C.

This is an example of a partition of S, which is generally a collection $B_1, B_2, \ldots$ of disjoint events (could be infinite) such that $\bigcup_i B_i = S$.

Theorem 8: If $B_1, B_2, \ldots$ is a partition of S with all $P(B_i) > 0$, then

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

Proof: …

Continuing with the HIV example, suppose we also know $P(A \mid B^c) = 0.005$ ("false positive").

We can now calculate $P(A)$.

useful expressions for calculation - III

Much to my amusement, Theorem 8 gets a grandiose title: "THE! LAW! OF! TOTAL! PROBABILITY!!!"

Now, in the HIV example, we also might be interested in $P(B|A)$, the chance of an HIV+ person testing positive.

A little algebra:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}$$

In our example this is $\frac{0.0021094}{0.0070988} = 0.2971.$

Bayes' rule

Theorem 9: If $B_1, B_2, \ldots$ is a partition of S with all $P(B_i) > 0$, then

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_i P(A \mid B_i)P(B_i)}$$

Proof:

independence

motivation - revisit the die toss example

I'll roll a six-sided die. $S = \{1, 2, 3, 4, 5, 6\}$. Consider these events:

$$\begin{aligned} A &= \{2, 5\}, \\ B &= \{2, 4, 6\} \end{aligned}$$

So $P(A) = \frac{2}{6} = \frac{1}{3}$.

What if I peek and tell you "Actually, B occurred". What is the probabality of A given this partial information? It is $\frac{1}{3}$.

The probability of A didn't change after the new information:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A)$$

definition(s) of independence

A and B are (pairwise) independent (notation $A \perp B$) if:

$$P(A \cap B) = P(A)P(B)$$

No requirement for $P(A)$ or $P(B)$ to be positive. In fact $\ldots$ see the suggested problems for Chapter 1.

$A_1, A_2, A_3, \ldots$ (possibly infinite) are (mutually) independent if for any finite subcollection of indices $I = \{i_1, \ldots, i_n\}$:

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

independence of two classes of events

Note that if $A \perp B$, then also $A \perp B^c$ and so on. Consider:

$$\begin{aligned} \mathcal{A} &= \{\emptyset, A, A^c, S\} \\ \mathcal{B} &= \{\emptyset, B, B^c, S\} \end{aligned}$$

Classes of events $\mathcal{A}$ and $\mathcal{B}$ are independent all pairs of events with one chosen from each class are independent.

The suggests a concept of "independent experiments", which will be revisited.

the "any" and "all" style of examples

(Note: in probability modeling, independence is usually assumed.)

A subway train is removed from service if any of its doors are stuck open. There is a probability p of a door getting stuck open on one day of operations. A train has n doors.

Example question: what is the chance a train is removed from service due to stuck doors on one day of operations?

p^n "all doors fail"

$1 - p^n$ "not all doors fail"

$(1 - p)^n$ "no doors fail"

$1 - (1 - p)^n$ "not no doors fail, in other words any doors fail" real valued functions with arguments that live inside sample spaces

the main focus of this course

We'll use "probability measure" throughout the course, but our main focus will be a different and equally strange object.

Recall that sample space is often arbitrary and difficult or impossible to describe.

It turns out usually we're ultimately interested in a real number that is associated with the random outcome, rather than the random outcome itself.

Consider a coin tossing game with $S = \{H, T\}$, which might be repeated, from which a multitude of examples can be invented $\ldots$.

Consider also the notion of picking a real number uniformly from $(0, 1)$ $\ldots$.

Eventually we will not even bother with an underlying S explicitly.

random variable

A random variable is a function whose domain is a sample space and whose range is $\mathbb{R}$.

Naming convention: Roman letters near the end of the alphabet X, Y, $X_1, X_2, \ldots$.

Another strange convention - almost always omit the function's "argument".

We will never draw a picture of a random variable, or compute a derivative or an integral of one.

We will instead focus on the defining property of a random variable: its distribution.

Perversely, we will lack the math to actually define distribution rigorously. Informally, the distribution of a random variable X is the rule that assigns probabilities to values of X.

assigning probabilities to values of X

As rigorous as we can get is mainly as follows.

A distribution is the rule that assigns probabilities that X takes on values in all intervals (closed, open, infinite, whatever), and simple set operations on intervals, e.g.

$$P(X \leq 1) \qquad P(\{X \leq 1\} \cap \{X \geq 1\}) = P(X = 1) \qquad P(X > -15)$$

The actual numbers above aren't important (1 and -15) and often generic statements are made using dummy placeholders like:

$$P(X \leq a) \qquad P(X = u) \qquad P(X > c)$$

With by far the most common being just "little x" as in $P(X = x)$ and $P(X \leq x)$.

complete descriptions of distributions, and other properties

If you know the distribution of a random variable, you know everything about it.

Most of the rest of this course will be occupied with:

different ways (all equivalent) to completely and uniquely describe distributions. These ways will always be functions (in this course from $\mathbb{R}$ to $\mathbb{R}$)

examples of random variables so important in practice that their distributions have special names.

examples of otherwise useless random variables useful as exercises

other not necessarily unique properties of distributions.

your first complete distribution descriptor

Suppose you have any random variable X. Its distribution can be completely described by the following function:

$$F_X(x) = P(X \leq x)$$

This is called the cumulative distribution function for X or cdf.

The subscript X is usually omitted unless required for clarity.

The domain is all of $\mathbb{R}$.

Note: that the cdf characterizes a distribution is actually a theorem which we lack the tools to prove.

defining properties of all cdf no matter what

Theorem: For any r.v. X, its cdf $F(x)$ has the following properties:

$$\lim_{x \to -\infty} F(x) = 0,$$

$$\lim_{x \to \infty} F(x) = 1,$$

and $F(x)$ is right-continuous, i.e.

$$\lim_{x \to a+} F(x) = F(a).$$

The proof of this theorem uses The Continuity Theorem and its corollary, and is left as an exercise.

(advanced note: any function with these properties is a cdf for some X)

discrete random variables

a large class of random variables

Discrete random variables take on a finite or countably ("list-able") set of real outcomes.

e.g. the coin toss game, and tossing a coin until the first head appears.

A more convenient complete distribution descriptor is the collection of probabilities of the set of outcomes, called the probability mass function or pmf:

$$p(x) = P(X = x)$$

This function is non-zero on the values of X, and formally 0 otherwise (usually just a formality).

pmf and cdf are "equivalent"

Theorem: for any discrete random variable X, the pmf and the cdf can be derived from each other.

Proof: next class

some important discrete random variables with special named distributions

the Bernoulli(p) distributions - fundamental building blocks

If a random variable takes on values 1 and 0 with probabilities p and $1 - p$ (for some fixed $0 < p < 1$), it is said to have a Bernoulli distribution with parameter p, or Bernoulli(p).

It doesn't really matter what the underlying sample space S actually is:

1. toss a die; $S = \{1, 2, 3, 4, 5, 6\}$; define $X_1(1) = X_1(2) = X_1(3) = 0$ and $X_1(4) = X_1(5) = X_1(6) = 1$
2. flip a coin; $S = \{H, T\}$; define $X_2(H) = 0$ and $X_2(T) = 1$

X_1 and X_2 have the same distribution, Bernoulli$\left(\frac{1}{2}\right)$.

Bernoulli(p) pmf and cdf

$$p(x) = \begin{cases} 1 - p & : \quad x = 0, \\ p & : \quad x = 1 \end{cases} \quad = \quad p^x (1 - p)^x \text{ for } x \in \{0, 1\}$$

$$F(x) = P(X \leq x) = \begin{cases} 0 & : \quad x < 0, \\ p & : \quad 0 \leq x < 1 \\ 1 & : \quad x \geq 1 \end{cases}$$

random variables - recap

the functions so far

1. Probabilitity measure: $P : \mathcal{A} \longrightarrow \mathbb{R}$ and satisfies the three axioms. In general no "picture" possible, because its domain is a collection of events.

2. Random variable $X : S \longrightarrow \mathbb{R}$. In general no "picture" possible, because its domain is a sample space. We care about: its distribution.

3. Cumulative distribution function F for the random variable X. Defined as $F(x) = P(X \leq x)$. A picture is possible, and does give some information of limited use.

recall "Example 1" "Example 2" and "Example 3"

Example 1: Toss a coin. $S = \{H, T\}$. $X = 1$ if H appears and $X = -1$ if T appears.

Example 2: Toss a coin repeatedly until H appears. $S = \{H, TH, TTH, \dots\}$. X_2 is the number of tosses required.

Example 3: Pick a real number "uniformly" between 0 and 1. $S = (0, 1)$. X_3 is the identity function. So:

$$F_{X_3}(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$

defining properties of a CDF

Theorem: For any r.v. X, its cdf $F(x)$ has the following properties:

$$\lim_{x \to -\infty} F(x) = 0,$$

$$\lim_{x \to \infty} F(x) = 1,$$

and $F(x)$ is right-continuous, i.e.

$$\lim_{x \to a+} F(x) = F(a).$$

The proof of this theorem uses The Continuity Theorem and its corollary. I'll do the middle property (see next three (new!) slides for an extra detail). The other two are left for you.

(advanced note: any function with these properties is a cdf for some X.)

notes — limits at infinity, and right-continuity (Part I)

There is the general definition of function limit, and function continuity. Here are some useful facts.

One can establish:

$$\lim_{x\to\infty} f(x) = L$$

by showing that $\lim_{n\to\infty} f(x_n) = L$ for any sequence $x_1 \le x_2 \le x_3 \le \cdots$ that goes to ∞.

notes — limits at infinity, and right-continuity (Part II)

One can also establish:

$$\lim_{x \to -\infty} f(x) = L$$

by showing that $\lim_{n\to\infty} f(x_n) = L$ for any sequence $x_1 \geq x_2 \geq x_3 \geq \cdots$ that goes to $-\infty$.

notes — limits at infinity, and right-continuity (Part III)

Finally, one can establish that a function is right-continuous at a by showing that

$$\lim_{n\to\infty} f(x_n) = f(a)$$

for any non-increasing sequence $x_1 \geq x_2 \geq x_3 \geq \cdots$ that converges to a.

not done *properly* in class - the proof of the middle property (part I)

First, note that I changed what used to be the previous one slide into three slides with more explicitly useful information that is needed to prove all three of the cdf properties.

Theorem: If F is a cdf, $\lim_{x \to \infty} F(x) = 1$.

Proof: First, see previous slides for a way to prove a "limit at infinity". (My first regret in class was not using this more specific result, and trying to be casual with the whole thing.)

Let $x_1 \leq x_2 \leq x_3 \leq \cdots$ be any unbounded non-decreasing sequence, so that it converges to $+\infty$.

Define the events $A_i = \{X \leq x_i\}$ (my second regret was not carefully enough defining actual increasing sequence of events of interest.).

not done *properly* in class - the proof of the middle property (part II)

Then $A_i \subset A_{i+1}$ and $\cup_{i=1}^{\infty} A_i$ is just the event $\{X \in \mathbb{R}\}$, which we can call A. (OK think informally that $A = \{-\infty < X < \infty\}$.

So by The Continuity Theorem we have $\lim_{n\to\infty} P(A_n) = P(A)$. But $P(A) = 1$, since X has to take on some value. And $P(A_n) = P(X \leq x_n) = F(x_n)$ by construction of A_n and the definition of F. So $\lim_{n\to\infty} F(x_n) = 1$.

Since all this is true no matter what unbounded non-decreasing sequence we had started with we conclude:

$$\lim_{x\to\infty} F(x) = 1.$$

discrete random variables

a large class of random variables

Discrete random variables take on a finite or countably ("list-able") set of real outcomes.

e.g. the coin toss game, and tossing a coin until the first head appears.

A more convenient complete distribution descriptor is the collection of probabilities of the set of outcomes, called the probability mass function or pmf:

$$p(x) = P(X = x)$$

This function is non-zero on the values of X, and formally 0 otherwise (usually just a formality).

defining properties of a pmf

A function $p(x)$ is a pmf if:

$$p(x) \geq 0$$

and

$$\sum_{\{x \,:\, p(x)>0\}} p(x) = 1$$

important concept: pmf and cdf are "equivalent"

Theorem: for any discrete random variable X, the pmf and the cdf can be derived from each other.

Proof: Surprisingly fussy, so don't worry about it. The pmf to cdf direction is easy:

$$F(x) \sum_{y \leq x} p(x)$$

The other direction isn't so easy. The idea (more important than the proof) is that you can recover the pmf from the cdf by noting that X takes on values exactly where the cdf jumps, and the sizes of those jumps are exactly the required probabilities.

some important discrete random variables with special named distributions

the Bernoulli(p) distributions - fundamental building blocks

If a random variable X takes on values 1 and 0 with probabilities p and $1 - p$ (for some fixed $0 < p < 1$), it is said to have a Bernoulli distribution with parameter p, or Bernoulli(p).

The phrase "X has a Bernoulli distribution with parameter p" will be abbreviated as:

$$X \sim \text{Bernoulli}(p)$$

"identically distributed"

It doesn't really matter what the underlying sample space S actually is:

1. toss a die; $S = \{1, 2, 3, 4, 5, 6\}$; define $X_1(1) = X_1(2) = X_1(3) = 0$ and $X_1(4) = X_1(5) = X_1(6) = 1$

2. flip a coin; $S = \{H, T\}$; define $X_2(H) = 0$ and $X_2(T) = 1$

X_1 and X_2 have the same distribution, Bernoulli$\left(\frac{1}{2}\right)$, but they are not the same function.

We say X_1 and X_2 are "identically distributed", and therefore the same as far as probability is concerned.

Bernoulli(p) pmf and cdf

$$p(x) = \begin{cases} 1-p & : \quad x = 0, \\ p & : \quad x = 1 \end{cases} \quad = \quad p^x(1-p)^{1-x} \text{ for } x \in \{0,1\}$$

$$F(x) = P(X \leq x) = \begin{cases} 0 & : \quad x < 0, \\ 1-p & : \quad 0 \leq x < 1 \\ 1 & : \quad x \geq 1 \end{cases}$$

Often used as a model for an "experiment" or other random process that either produces an event A of interest, or it doesn't. If A is some event we can define the useful indicator function:

$$I_A = \begin{cases} 1 & \text{when } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

I_A is a random variable. It will have a Bernouilli distribution with parameter $p = P(A)$.

thank you for subscribing to Bernoulli facts!

Sometimes we'll say q instead of $1 - p$.

It's not clear why the outcomes 0 and 1 are important and useful. Why not just focus on the underlying random outcome? More later.

It starts with the idea of Bernoulli process, which I'll introduce now but cannot completely describe until later.

Bernoulli process

Consider an experiment with an event of interest A. Sometimes (unfortunately!), when A occurs we call it a "success". Otherwise, it's a "failure".

Suppose we replicate the experiment. Maybe a finite number n times, or maybe indefinitely.

Each experiment repetition is independent of all the others. (What does this mean?)

If we let the result of the i^{th} replication be $X_i = I_A$, then $X_1, X_2, \ldots$ is called a Bernoulli process (or "sequence of Bernoulli trials").

the Binomial distributions

Stop a Bernoulli(p) process after n trials. Count the number of "successes", or 1's.

This is a random variable. Call it X.

What is the p.m.f. of X?

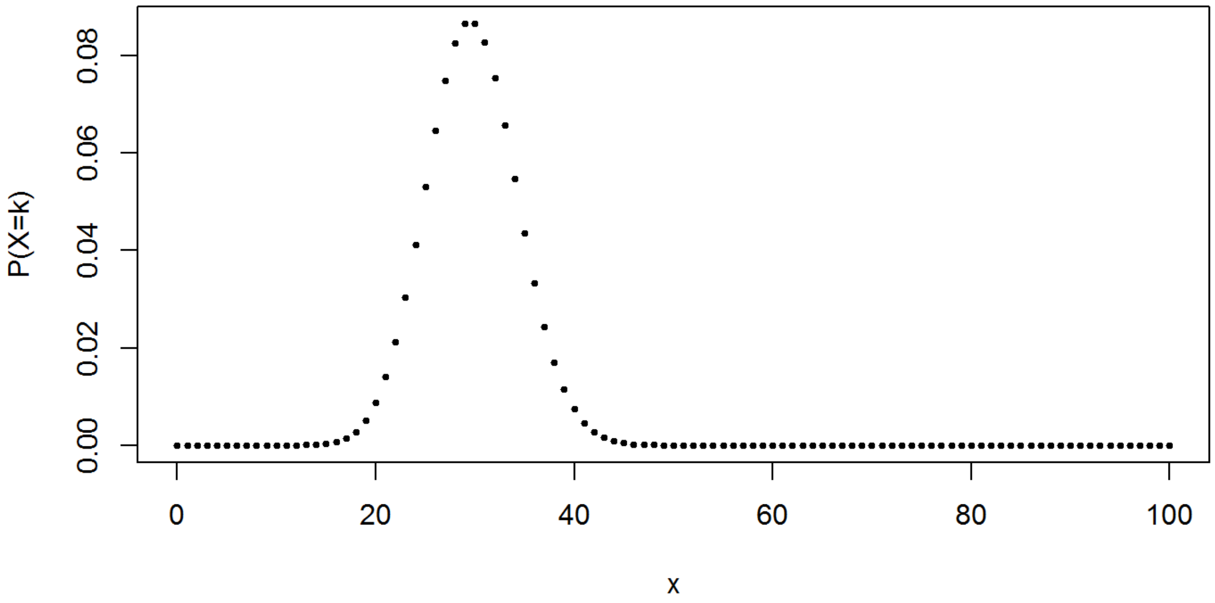
$$p(k) = P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & x \in \{0, \dots, n\} \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid pmf? Yes. ("Binomial Theorem")

We say X has a Binomial(n, p) distribution, or $X \sim \text{Binomial}(n, p)$.

a few examples…

The probability that someone is HIV+ given that their ELISA test comes back positive is 0.2971. Suppose we have 100 people with a positive ELISA test. How might one visualize the distribution of the number of people who are HIV+?



…a few examples

The extreme cases have already been considered.

$$P(X = n)$$

$$P(X = 0)$$

$$1 - P(X = n)$$

$$1 - P(X = 0)$$

the geometric distributions

Consider a Bernoulli(p) process. Count the number of trials until the first "success".

This is a random variable. Call it X.

What is the p.m.f. of X?

$$p(k) = P(X = k) = \begin{cases} (1-p)^{k-1}p & x \in \{1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid pmf? Yes.

We say X has a geometric distribution with paramter p, or $X \sim \text{Geometric}(p)$.

the "negative binomial" distributions

Consider a Bernoulli(p) process. Count the number of trials until the r^{th} "success".

This is a random variable. Call it X.

What is the p.m.f. of X?

$$p(k) = P(X = k) = \begin{cases} \binom{k-1}{r-1} p^k (1-p)^{k-r} & x \in \{r, r+1, r+2, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid pmf? Yes (a little obscure to figure out)

We say X has a negative binomial distribution with paramters p and r, or $X \sim \text{NegBin}(p, r)$.

recap

the functions so far

1. Probabilitity measure: $P : \mathcal{A} \longrightarrow \mathbb{R}$ and satisfies the three axioms. In general no "picture" possible, because its domain is a collection of events.

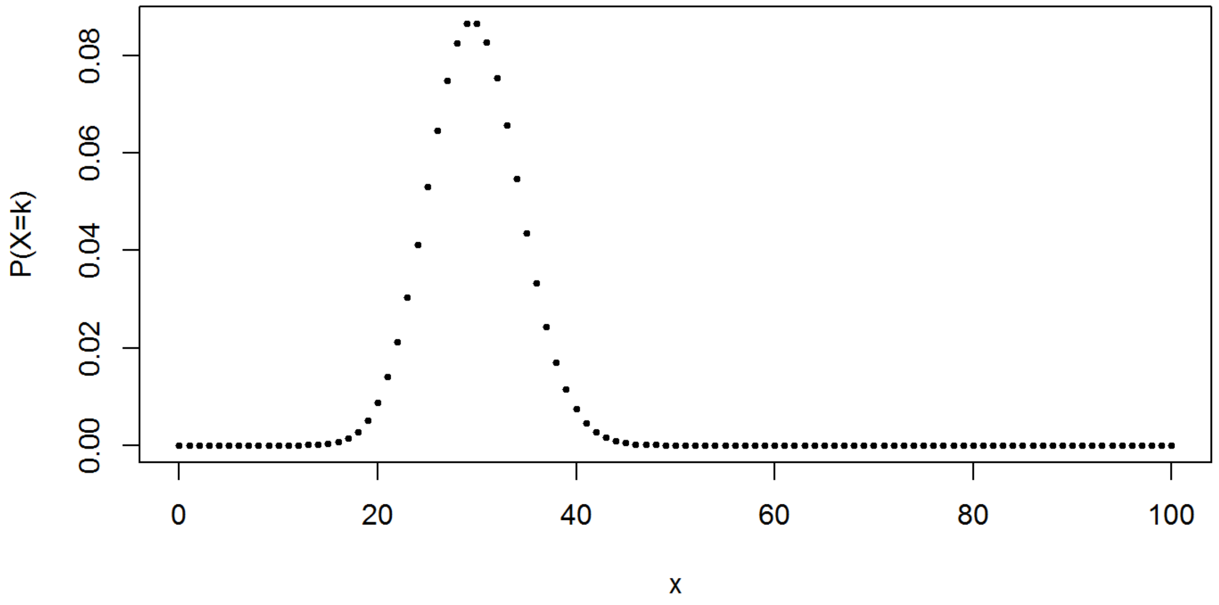
2. Random variable $X : S \longrightarrow \mathbb{R}$. In general no "picture" possible, because its domain is a sample space. We care about: its distribution.

3. Cumulative distribution function F for the random variable X. Defined as $F(x) = P(X \leq x)$. Completely characterizes a distribution. A picture is possible, and the picture does give some information of limited use. NEW: Has a few technical properties of interest.

4. For a discrete random variable, there is also a probability mass function $p(x) = P(X = x)$. Completely characterizes a distribution. A picture is possible, and the picture can be informative. Has a few technical properties of interest.

Binomial(n, p) distributions

The probability that someone is HIV+ given that their ELISA test comes back positive is 0.2971. Suppose we have 100 people with a positive ELISA test. How might one visualize the distribution of the number of people who are HIV+?



…a few more examples

The extreme cases have already been considered.

$$P(X = n)$$

$$P(X = 0)$$

$$1 - P(X = n)$$

$$1 - P(X = 0)$$

Problem solving hints: look for cases where the number of trials is fixed and one is interesting in counting occurences of something.

the geometric distributions

Consider a Bernoulli(p) process. Count the number of trials until the first "success".

This is a random variable. Call it X.

What is the p.m.f. of X?

$$p(k) = P(X = k) = \begin{cases} (1-p)^{k-1}p & k \in \{1, 2, 3, \ldots\} \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid pmf? Yes.

We say X has a geometric distribution with paramter p, or $X \sim \text{Geometric}(p)$.

Interesting property: "memoryless"

the "negative binomial" distributions

Consider a Bernoulli(p) process. Count the number of trials until the r^{th} "success".

This is a random variable. Call it X.

What is the p.m.f. of X?

$$p(k) = P(X = k) = \begin{cases} \binom{k-1}{r-1} p^k (1-p)^{k-r} & x \in \{r, r+1, r+2, \ldots\} \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid pmf? Yes (a little obscure to figure out)

We say X has a negative binomial distribution with paramters p and r, or $X \sim \mathrm{NegBin}(p, r)$.

the "negative binomial" distributions

Consider a Bernoulli(p) process. Count the number of trials until the r^{th} "success".

This is a random variable. Call it X.

What is the p.m.f. of X?

$$p(k) = P(X = k) = \begin{cases} \binom{k-1}{r-1} p^r (1-p)^{k-r} & k \in \{r, r+1, r+2, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid pmf? Yes (a little obscure to figure out)

We say X has a negative binomial distribution with paramters p and r, or $X \sim \text{NegBin}(p, r)$.

added after class - I

There was an error in what I put on the board. As penance, I'll give you some more information you might find useful in your travels, but is not so critical for this course.

First, the formula $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ has a few conventions and extensions associated with it. That version of the formula is the most direct translation of its combinatoric interpretation (number of ways to pick k out of n.) By convention, because of the combinatoric interpretation, when $n = 0$ we say $\binom{n}{k} = 0$. There is "no way" to pick k items out of 0.

But that formula is often inconvenient for calculation. So often one uses (in fact we'll do this on October 12) this version:

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

added after class - II

Note the numerator has k terms in it. The nice thing about this formula is that it can be extended to negative numbers (and indeed to any real number, which we won't bother with) as follows:

$$\binom{-n}{k} = \frac{(-n)(-n-1)(-n-2)\cdots(-n-k+1)}{k!}$$

By convention $\binom{-n}{0} = 1$. One can show (easy to find on internet) a version of the Binomial Theorem with negative exponent:

$$(1+x)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k, \text{ for } |x| < 1$$

It is expressed this way rather than something like $(a+b)^{-n}$ to avoid difficulties in saying where the infinite series converges.

added after class - III

The eventual goal here is to prove the NegBin(r, p) sums to 1. But that pmf has $\binom{r-1}{k-1}$ in it, which isn't exactly what appears in this "negative" Binomial theorem.

The following will turn out to be useful (to follow along note there are k terms on top):

$$\begin{aligned}\binom{k+r-1}{k} &= \frac{(k+r-1)(k+r-2)\cdots r}{k!} \\ &= (-1)^k \frac{(-r-(k-1))(-r-(k-2))\cdots(-r)}{k!} \\ &= (-1)^k \binom{-r}{k}\end{aligned}$$

Also we'll use $\binom{k-1}{r-1} = \binom{k-1}{k-r}$ (think combinatorically to see this is true.)

added after class - IV

Now we'll get back to the NegBin(r, p) p.m.f.:

$$\begin{aligned}
\sum_{k=r}^{\infty} p(k) &= \sum_{k=r}^{\infty} \binom{k-1}{r-1} p^r (1-p)^{k-r} \\
&= \sum_{k=r}^{\infty} \binom{k-1}{k-r} p^r (1-p)^{k-r} \\
&= \sum_{k=0}^{\infty} \binom{k+r-1}{k} p^r (1-p)^{k} \\
&= p^r \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} (1-p)^{k}
\end{aligned}$$

added after class - V

Continued:

$$\begin{aligned}\sum_{k=r}^{\infty} p(k) &= p^r \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} (1-p)^k \\ &= p^r \sum_{k=0}^{\infty} \binom{-r}{k} (-1+p)^k \\ &= p^r (1-1+p)^{-r} = 1\end{aligned}$$

All these developments are not hard to find in books and elsewhere. The value is in the journey more so than the final destination. We might need something from all this later on, or maybe not. You will not be responsible for it in this course.

Also, note that the (only) reason this distribution is called "negative binomial" is because of the use of the "negative" version of the Binomial Theorem.

the "hypergeometric distributions"

Many examples from Ch. 1 (quality control, Lotto, some of the balls in urns etc.)

Consider a Bernoulli process, stopped after n trials in which there were r "successes". Let X be the number of successes in the "first" m out of n trials. Note: this relationship between hypergeometric and Bernoulli process is correct, but a bit contrived. Best to think of it in terms of sampling without replacement from finite populations. Statistics students will find they don't encounter this distribution all that often.

What is the p.m.f. of X?

$$p(k) = P(X = k) = \begin{cases} \dfrac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}} & : k \in \{0, \ldots, r\} \\ 0 & : \text{ otherwise} \end{cases}$$

Is this a valid pmf? Yes. (Problem 34 from Chapter 1)

We say X has a hypergeometric distribution with paramters r, n, and m.

added after class - hypergeometric vs. binomial - I

You are not responsible for this material.

In Ch. 1 there were some urn and "quality control" questions that dealt with the as-yet unnamed hypergeometric distibutions.

Key to the hypergeometric distributions is the notion of "sampling without replacement".

For example, the urn has r black balls and $n - r$ white balls. If you select a ball, it will be black with probability $p = r/n$ and white with $1 - p = (n - r)/n$.

Now let's say you put the ball back in the urn, and then select a ball again. "Sampling with replacement." The black/white probabilities don't change. Successive draws are independent. If you do this m times, the total number of black balls selected will have a Binomial(m, p) distribution.

added after class - hypergeom vs. binom - II

(You are not responsible for this material.)

If you don't put the ball back in the urn after each selection you have "Sampling without replacement." Successive draws are not independent, because the probability of choosing a black ball in in the second selection depends on the outcome of the first selection. If you select m balls the total number of black balls selected will have a hypergeometric distribution with parameters n, r, and m.

Let's look at the lack of independence in more detail. If $n = 100$ and $r = 50$, then the chance of a black ball is $50/100$ on the first select and then either $49/99$ or $50/99$ on the second. The probability changes - lack of independence - but it doesn't change by that much really.

added after class - hypergeom vs. binom - III

(You are not responsible for this material.)

Increase n to 10000 and r to 5000 (keeping r/n the same.) Now the probabilities are $5000/10000$ and $4999/9999$ or $5000/9999$. Still different, but now even closer. In some sense, the lack of independence is becoming smaller.

There is a way that the hypergeometric "converges" to the binomial, if you allow n and r to increase in a way that r/n converges to a constant. We are going to look at the following:

$$\lim_{\substack{n\to\infty \\ r\to\infty \\ r/n\to p}} \frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}$$

added after class - hypergeom vs. binom - IV

(You are not responsible for this material.)

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

$$= \frac{m!}{k!(m-k)!}\,\frac{\overbrace{r(r-1)\cdots(r-k+1)}^{k\text{ terms}}\ \overbrace{(n-r)(n-r-1)\cdots(n-r-(m-k)+1)}^{m-k\text{ terms}}}{\underbrace{n(n-1)\cdots(n-m+1)}_{m\text{ terms}}}$$

$$= \binom{m}{k}\underbrace{\frac{r}{n}\,\frac{r-1}{n-1}\cdots\frac{r-k+1}{n-k+1}}_{k\text{ terms}}\ \underbrace{\frac{n-r}{n-k}\,\frac{n-r-1}{n-k-1}\cdots\frac{n-r-(m-k)+1}{n}}_{m-k\text{ terms}}$$

added after class - hypergeom vs. binom - V

(You are not responsible for this material.)

Now let n and r go to ∞, but in proportion so that r/n goes to a constant $0 < p < 1$.

There is a fixed number of terms, so the limit can be applied one term at a time. The first k terms convege to $r/n = p$ and the last $m - k$ terms converge to $(n - r)/n = 1 - p$. So the whole thing converges to:

$$\binom{m}{k} p^k (1 - p)^{m-k}$$

digression — constants as "random variables"

In calculus etc. you may have (unconsciously) considered things like $f(x) + a$ for real constant a. This could have a few interpretations, such as:

1. The sum of the numbers $f(x)$ and a.
2. The value of the function $f + g$ evaluated at x, in which it happens that $g(x) = a$ for all x.

We do this in probability as well. We can allow a random variable X to be some constant a no matter what. Then X is discrete with :

$$p(x) = P(X = x) = \begin{cases} 1 & : \ x = a \\ 0 & : \ \text{otherwise}, \end{cases}$$

$$F(x) = P(X \le x) = \begin{cases} 0 & : \ x < a \\ 1 & : \ x \ge a. \end{cases}$$

the Poisson distributions - I

Named after a French guy called Poisson.

Can be defined completely abstractly just by declaring X has a Poisson distribution with parameter λ, or X ∼Poisson(λ), if X has p.m.f:

$$p(k) = P(X = k) = \begin{cases} \frac{\lambda^k e^{-\lambda}}{k!} & : k \in \{0, 1, 2, \dots\} \\ 0 & : \text{otherwise.} \end{cases}$$

Is this a valid p.m.f.? Yes.

But this does not come close to explaining the importance of the Poisson distributions.

Bernoulli process — with a time scale

Consider a Binomial(n, p) distribution. Let's introduce a time scale to the underlying Bernoulli(p) process.

Step 1. Take a fixed time interval $(0, t)$ and divide it into $n_1 = n$ subintervals and let $p_1 = p$. Let's say one Bernoulli(p_1) trial happens inside each subinterval, and we keep track of the number of "successes".

A few "general" observations:

Only one success can happen inside each subinterval. Results in non-overlapping collections of subintervals are independent. Successes happen at a "rate" of about $n \cdot p$ (intuitively) over the whole time $(0, t)$, and the number of k successes has a Binomial(n_1, p_1) distribution. t didn't matter — if we double it to $2t$ the rate simply doubles also, to $2np$.

double the number of intervals

Step 2. Divide the same interval into $n_2 = 2n$ subintervals.

We want a trial to happen inside each subinterval, but we want the same overall "rate" of success.

So now we allow a Bernoulli(p_2) trial with $p_2 = p/2$ to occur inside each subinterval.

The same "general" observations continue to hold, except now the number of successes has a Binomial$(2n, p/2)$ distribution.

Step 4. Double the number of intervals again… (so now $(0, t)$ has $4n$ intervals with a Bernoulli($p/4$) trial in each)…

pass to the limit

Define λt ("rate") to be fixed and always equal to np. This implies $p = \frac{\lambda t}{n}$.

What happens to Binomial$\left(n, \frac{\lambda t}{n}\right)$ distributions as $n \to \infty$?

λ is the rate of occurrences per unit time.

Examples and more discussion next time.

binomial to Poisson

From last time:

Take an interval $(0, t)$ and an $np = \lambda t$;

Split into subintervals;

Stick a Bernoulli trial in each subinterval;

In such a way that you'd expect about the same number λt of events to occur in $(0, t)$.

Observe that only one event can happen in each subinterval and non-overlapping collections of subintervals are independent.

Add for today:

Rather than $(0, t)$ we could have taken any arbitrary (s, t) (different t!). We'll expect about $\lambda(t - s)$ events to take place within (s, t).

pass to the limit

Define $\lambda(t-s)$ be fixed and always equal to np. This implies $p = \frac{\lambda(t-s)}{n}$.

What happens to $X_n \sim \text{Binomial}\left(n, \frac{\lambda(t-s)}{n}\right)$ distributions as $n \to \infty$?

$$\lim_{n\to\infty} P(X_n = k) = \frac{[\lambda(t-s)]^k}{k!} e^{-\lambda(t-s)}$$

Note: λ itself is the rate of occurrences per unit time.

Let $Y \sim \text{Poisson}(\lambda(t-s))$. The above implies more:

$$\lim_{n\to\infty} P(X_n \in A) = P(Y \in A),$$

for $A \subset \mathbb{R}$. So the "distributions" converge.

more on the limit result

This limit result serves a few purposes:

1. Because of the speed and accuracy of the convergence, one can approximate binomial probabilities.

2. Motivate use of Poisson distribution as a suitable probability model (large n, small p, one-at-a-time events, etc.)

binomial approximation example

Suppose $X \sim \text{Binomial}(n, p)$ with $n = 10000$ and $p = 0.001$.

Calculate and approximate $P(X \leq k)$ for $k \in \{0, 1, 2, 3, 4, 5\}$.

Calculation: $\sum_{i=0}^{k} \binom{10000}{i} (0.001)^i (0.999)^{10000-i}$

Approximation: $\sum_{i=0}^{k} \frac{10^i e^{-10}}{i!}$, using $\lambda = np = 10$

	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Binom	4.517E-05	4.522E-04	2.263E-03	7.549E-03	1.889E-02	3.780E-02
Poisson	4.540E-05	4.540E-04	2.270E-03	7.567E-03	1.892E-02	3.783E-02
Diff	-2.266E-07	-1.814E-06	-7.031E-06	-1.740E-05	-3.027E-05	-3.785E-05

the "completely random" nature of the Bernoulli process

Only one "event" can happen at a time.

Non-overlapping sets of trials are independent.

Memoryless (as seen via Geometric distributions).

Another idea: suppose by the n^{th} trial exactly 1 event has occurred. Let Y be the index of the trial where the event occurred. This is a random variable. What is it's distribution?

Note: sometimes Y is said to have a "discrete uniform distribution on $\{1, 2, \ldots, n\}$"

the Poisson process - I

Suppose as time unfolds, events occur, and we keep track of the number of events that occur over time. Denote by $N(t)$ the number of events that happen inside $(0, t]$.

(So $N(t) - N(s)$ is the number of events that happen in some $(s, t]$ with $s < t$).

Let's say we want the occurrence of events to be "completely random" in analogy to a Bernoulli process. Let's specify the following:

Only one event at a time.

Counts of events that happen in non-overlapping intervals are independent.

the Poisson process - II

Then if we add a contant "rate" λ per unit time of occurrences it turns out it must be that:

$$P(N(t) - N(s) = k) = \frac{[\lambda(t-s)]^k}{k!} e^{-\lambda(t-s)}.$$

We'll say $N(0) = 0$.

Many generalizations exist. The Poisson processes occupy a central role in probability.

practical examples (stolen from Schay)

Customers enter a store at a rate of 1 per minute. Find the probabilities that:

1. More than one will enter in the first minute.
2. More than two will enter in the first two minutes.
3. More than one will enter in each of the first two minutes.

Why and why not might a Poisson process model be suitable here?

not done in lecture: Chapter 2 Question 31 - Two Ways

method 1 - alter the rate by the suggested time fraction

"Phone calls are received at a certain residence as a Poisson process with parameter $\lambda = 2$ per hour."

"a. If Diane takes a 10-min. shower, what is the probability that the phone rings during that time."

10 minutes is 1/6 of an hour. So the relevant rate parameter for a Poisson distribution is $2 \cdot 1/6$. Let X be the number of phone calls received during the shower, so that $X \sim \text{Poisson}(2/6)$. We want the probability of any phone calls $P(X > 0)$, which is $1 - P(X = 0) = 1 - e^{-2/6} = 0.2834687$

method 1 - alter the rate by the suggested time fraction

"b. How long can her shower be if she wishes the probability of receiving no phone calls to be at most .5?"

The rate of calls is 2 per hour. We want the fraction, say θ, of one hour to be such that $P(X = 0) = 0.5$.

$$\begin{aligned}
\frac{(2\theta)^0 e^{-2\theta}}{0!} &= 0.5 \\
e^{-2\theta} &= 0.5 \\
-2\theta &= \log(0.5) \\
\theta &= -\frac{\log(0.5)}{2} = \frac{\log(2)}{2} = 0.3465736
\end{aligned}$$

The shower can be 20.7944154 minutes long.

method 2 - Poisson process method

"a. If Diane takes a 10-min. shower, what is the probability that the phone rings during that time."

some simplified notation

It is common to deal with probabilities of intersections of events relating to random variables.

$$P(\{X \leq 3\} \cap \{Y > 5\})$$

Usually we omit the braces and use a comma:

$$P(X \leq 3,\ Y > 5)$$

recap

A Poisson process $N(t)$ counts the number of event that happen between "time" 0 and "time" t.

The rate per unit time is λ.

Probability of k events between times s and t are:

$$P(N(t) - N(s) = k) = \frac{[\lambda(t-s)]^k e^{-\lambda(t-s)}}{k!}$$

and $N(0) = 0$ always, so when dealing with intervals $(0, t]$ we can just use

$$P(N(t) = k) = \frac{[\lambda t]^k e^{-\lambda t}}{k!}.$$

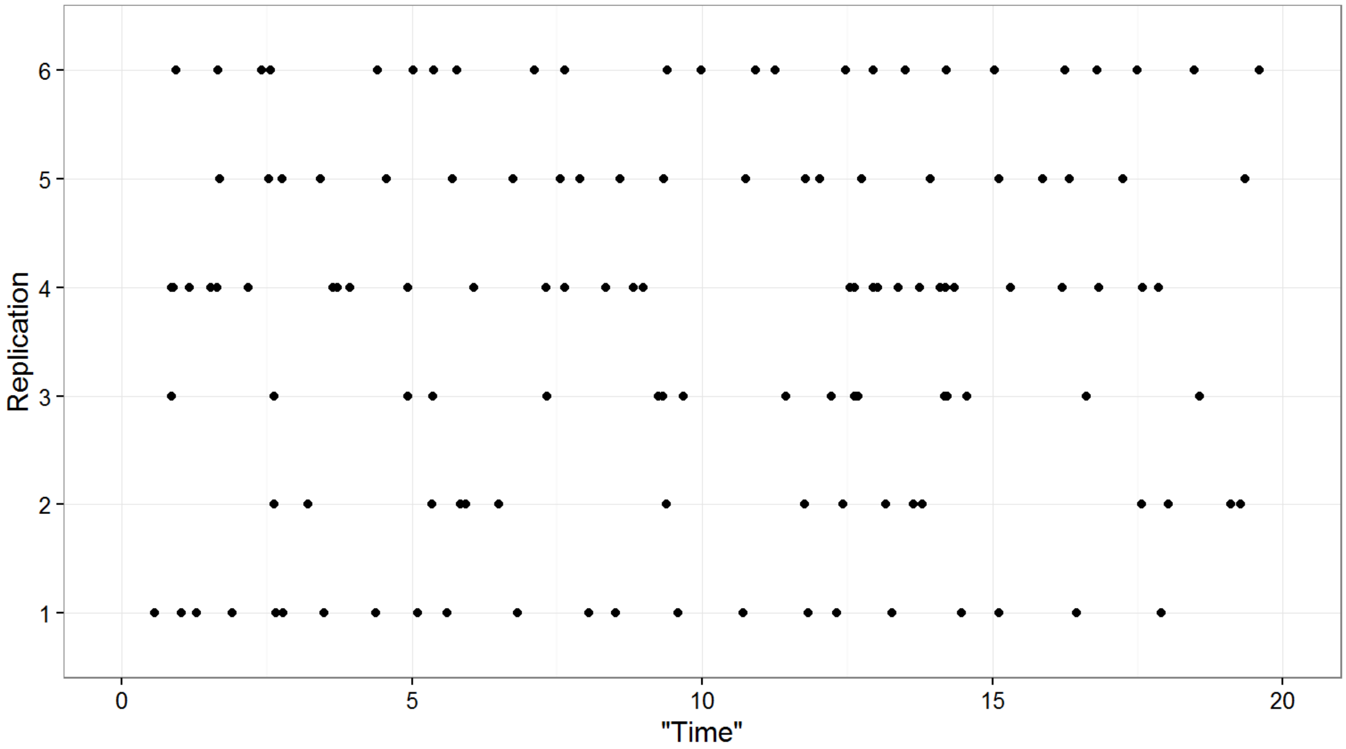
practical examples (stolen from Schay)

Q7.1.1 (p. 236) Customers enter a store at a rate of 1 per minute. Find the probabilities that:

1. More than one will enter in the first minute. (from last time: 0.264)
2. More than two will enter in the first two minutes. (0.323)
3. More than one will enter in each of the first two minutes. $\left(0.264^2\right)$

DISASTROUS TREND SHOCKER PANIC HEADLINE

which ones are "completely random" (Poisson?)



Answer: 2, 3, 4

continuous random variables

from counting to measuring

Discrete random variables count things.

We need random variables to measure things.

Let X be such a random variable. As always it will have a cdf $F(x) = P(X \leq x)$, which will turn out to be continuous.

Main focus as always is on distribution, i.e. the collection of $P(X \in A)$ for $A \subset \mathbb{R}$.

Let's just worry about intervals and consider things like:

$$P(X \in (a, b]) = P(a < X \leq b) = F(b) - F(a).$$

density

If there is a ("piecewise") continuous function f such that:

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)\,dx$$

then we say X is "(absolutely) continuous" and has f as its probability density function (or pdf, or just density).

Note: a and b can be $-\infty$ or ∞.

density functions, meanings, consequences - I

Theorem: A pdf completely characterizes a distribution.

Proof: …

Some corollaries: the cdf F is continuous, $f \geq 0$ (wherever it is continuous), and $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Theorem: Any piecewise continuous, non-negative f with $\int f = 1$ can be a density function.

Proof: Too hard…

density functions, meanings, consequences - II

Advice: Always think of a density as living inside its integral.

Heuristic meaning of $f(x)$ can be:

$$f(x)\Delta x \approx \int_x^{x+\Delta x} f(x)\,dx = P(X \in (x, x+\Delta x]).$$

Pictures of densities can be useful, to show relative differences in probabilities.

density functions, meanings, consequences - III

If X is continuous, then for all $a \in \mathbb{R}$:

$$P(X = a) = P(X \in (a, a]) = \int_a^a f(x)\, dx = 0$$

Recall from the beginning "pick a random number in (0,1)" and its associated oddities.

For continuous random variables we have these conveniences:

$$P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b), \text{ etc.}$$

example

Pick a number "uniformly" from $(0, 1)$ and let X simply be that number. We have:

$$F(x) = P(X \leq x) = \begin{cases} 0 & : x < 0 \\ x & : 0 \leq x < 1 \\ 1 & : x \geq 1 \end{cases}$$

To find the density, just differentiate:

$$f(x) = F'(x) = \begin{cases} 1 & : 0 \leq x < 1 \\ 0 & : \text{otherwise} \end{cases}$$

The endpoints don't matter. We say X has a uniform distribution with parameters 0 and 1, or $X \sim \text{Uniform}[0, 1]$.

the $\text{Uniform}[a, b]$ distributions

Nothing special about 0 and 1. Pick a number between a and b and call it X. The density and cdf will be:

$$f(x) = \begin{cases} \frac{1}{b-a} & : x \in [a, b] \\ 0 & : \text{otherwise}; \end{cases}$$

$$F(x) = \begin{cases} 0 & : x < a \\ \frac{x-a}{b-a} & : x \in [a, b] \\ 1 & : x > b. \end{cases}$$

The density is used to calculate probabilities. Say $X \sim \text{Uniform}[-1, 2]$. We can calculate things like: $P(0 < X < 3/2)$, $P(-3 < X \leq 0)$, etc.

an (apparently) artificial example

Suppose f is defined as:

$$f(x) = \begin{cases} cx & : 0 < x \leq 1, \\ c(2-x) & : 1 < x \leq 2, \\ 0 & : \text{ otherwise.} \end{cases}$$

Determine c that makes f a valid density.

If X has f as its density, determine its cdf and calculate $P(0.75 < X < 1.5)$.

time to first event of a Poisson process

Let's say we have a Poisson process $N(t)$ with rate λ. The time of the first event is random. Call this time X.

What can we say about X? Can we completely describe its distribution?

Yes, because $F(x) = 1 - P(X > x)$ and $\{X > x\}$ is exactly equivalent to $\{N(x) \le 0\}$, so we can derive the cdf for X.

$$F(x) = P(X \le x) = \begin{cases} 0 & : x \le 0 \\ 1 - e^{-\lambda x} & : x > 0 \end{cases}.$$

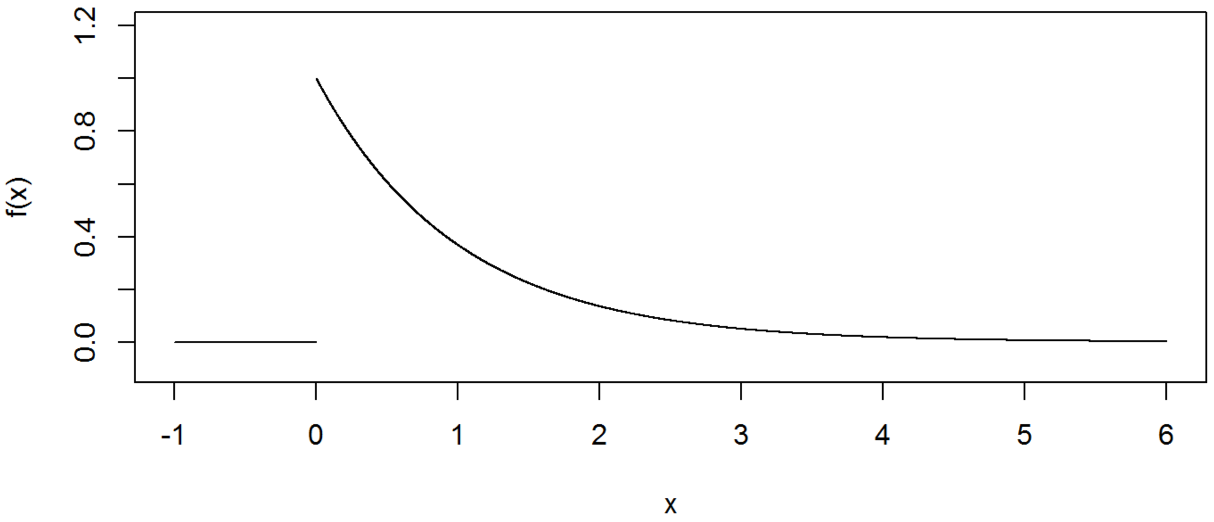
So the density is:

$$f(x) = F'(x) = \begin{cases} \lambda e^{-\lambda x} & : x > 0 \\ 0 & : \text{otherwise.} \end{cases}$$

the exponential distributions

In this case we say X has an exponential distribution with (rate) parameter λ, or $X \sim \mathrm{Exp}(\lambda)$.

Free picture for Exp(1) density:



what should we expect of a Poisson waiting time?

The Poisson process is the continuous time analogy of the Bernoulli process. Both intended to be "completely random" (process is memoryless and counts over disjoint intervals are independent.)

The exponential distributions, like the geometric, turn out to be memoryless.

But wait! There's more!

Theorem: The exponential distributions are the only continuous, memoryless distributions.

Proof: …

exponential example

The exponential distributions are commonly used as a model of "completely random failure". Examples include complex systems, electronic devices, and many others.

Suppose a haul truck diesel engine has a failure time X modeled using an exponential distribution. The rate is 1 failure every 5 years.

What is the probability that an engine will survive more than 3 years?

What is the probability that, out of a fleet of $n = 20$ engines, more than half will survive more than 3 years? (Assume independent failures.)

time to n^{th} event of a Poisson process

Let's say we have a Poisson process $N(t)$ with rate λ. The time of the ~~first~~ n^{th} event is random. Call this time X.

What can we say about X? Can we completely describe its distribution?

Yes, because $F(x) = 1 - P(X > x)$, and $\{X > x\}$ is exactly equivalent to $\{N(x) \leq n - 1\}$, so we can derive the cdf for X.

$$F(x) = P(X \leq x) = \begin{cases} 0 & : x \leq 0 \\ 1 - \sum_{i=0}^{n-1} \frac{[\lambda x]^i}{i!} e^{-\lambda x} & : x > 0 \end{cases}.$$

So the density is (after some fussy work):

$$f(x) = F'(x) = \begin{cases} \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} & : x > 0 \\ 0 & : \text{otherwise.} \end{cases}$$

generalization: the gamma distributions

Definition: the gamma function is defined as:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du, \qquad \alpha > 0.$$

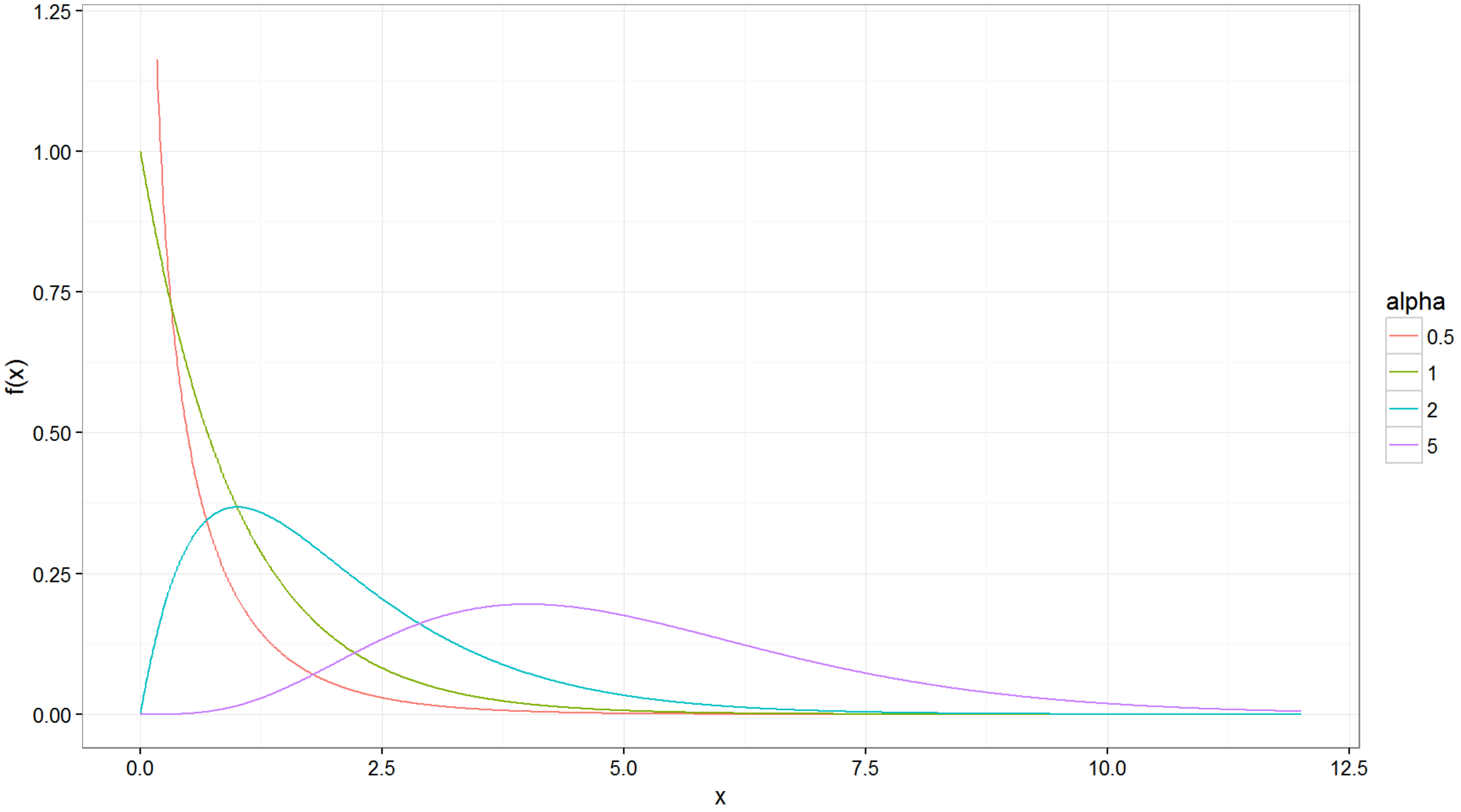
Many interesting properties, including $\Gamma(n) = (n-1)!$ for integer $n \geq 1$ (exercise 49 in book.)

The following function is a valid density for $\alpha > 0$ and $\lambda > 0$:

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & : x > 0 \\ 0 & : \text{ otherwise.} \end{cases}$$

Proof: …

free pictures of some Gamma(α, 1) densities



gamma distribution trivia

We can say X has a gamma distribution with "shape parameter"" α and "rate parameter" λ, or $X \sim \text{Gamma}(\alpha, \lambda)$.

Lots of things seem to (empirically) have gamma distributions: insurance claim amounts, crack growth models, earthquake times, neuron spike, etc.

Special cases: $\alpha = 1$ is exponential.

$\alpha = n$ positive integer gives the waiting time until the nth event in a Poisson process (also called "Erlang(n, λ)")

$\alpha = n/2$ and $\lambda = 1/2$ is called χ_n^2 and has applications in statistics. (More later.)

another fun Poisson process fact

Suppose at some fixed time t of a Poisson process we know $N(t) = 1$. In other words, exactly one "event" occured at some time before t.

This occurrence time is a random variable. Call it X. What is its distribution?

Let's try to derive its cdf:

$$F(x) = P(X \leq x) = \begin{cases} 0 & : x < 0 \\ ??? & : 0 \leq x \leq t \\ 1 & : x \geq t \end{cases}$$

exercise

Suppose at some fixed time $t > 0$ of a Poisson process we know $N(t) = n$.

Fix another time s with $0 < s < t$.

Find the distribution of the number X of events that occurred inside $[0, s]$.

(not done in class) n^{th} event density fussy work

$$
\begin{aligned}
\frac{d}{dx}\left(1-\sum_{i=0}^{n-1}\frac{[\lambda x]^i}{i!}e^{-\lambda x}\right) &= -\frac{d}{dx}\left(e^{-\lambda x}\sum_{i=0}^{n-1}\frac{[\lambda x]^i}{i!}\right)\\
&= -\left(-\lambda e^{-\lambda x}\sum_{i=0}^{n-1}\frac{[\lambda x]^i}{i!}+e^{-\lambda x}\sum_{i=1}^{n-1}\frac{\lambda[\lambda x]^{i-1}}{(i-1)!}\right)\\
&= \lambda e^{-\lambda x}\left(\sum_{i=0}^{n-1}\frac{[\lambda x]^i}{i!}-\sum_{i=1}^{n-1}\frac{[\lambda x]^{i-1}}{(i-1)!}\right)\\
&= \lambda e^{-\lambda x}\left(\sum_{i=0}^{n-1}\frac{[\lambda x]^i}{i!}-\sum_{i=0}^{n-2}\frac{[\lambda x]^{i}}{(i)!}\right)\\
&= \frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}
\end{aligned}
$$

(more) gamma (distribution) trivia

Recall:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du$$

Also recall density of Gamma(α, λ) distributions:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

This would be used to compute probabilities. $f(x)$ does not have a closed-form anti-derivative (as usual), so a numerical routine is used (as usual).

$\alpha = n/2$ and $\lambda = 1/2$ is called χ^2_n and has applications in statistics. (More today.)

another fun Poisson process fact

Suppose that at some fixed time t of a Poisson process we know $N(t) = 1$. In other words, exactly one "event" occured at some time before t.

This occurrence time is a random variable. Call it X. What is its distribution?

Let's try to derive its cdf:

$$F(x) = P(X \leq x) = \begin{cases} 0 & : x < 0 \\ ??? & : 0 \leq x \leq t \\ 1 & : x \geq t \end{cases}$$

exercise

Suppose that at some fixed time $t > 0$ of a Poisson process, we know $N(t) = n$, but we don't know when the n events occurred.

Fix another time s with $0 < s < t$.

Find the distribution of the number X of events that occurred inside $[0, s]$.

"Normal" distributions

A (continuous) random variable X has a "Normal" (or "Gaussian") distribution if its density is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Normal distributions are indexed by two parameters: the "mean" $-\infty < \mu < \infty$ and the "standard deviation" $\sigma > 0$.

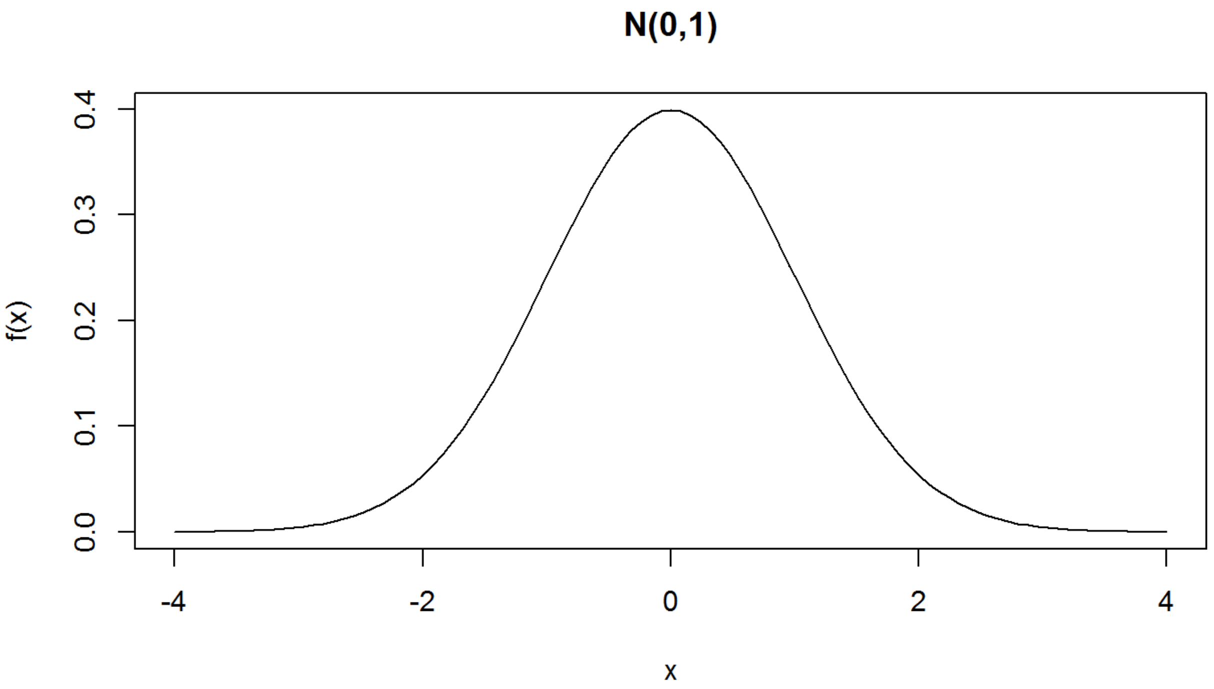
It is also common to index by μ and $\sigma^2 = (\sigma)^2$ called the "variance".

This is called a "reparametrization".

Common notation: $X \sim N(\mu, \sigma)$ or $X \sim N(\mu, \sigma^2)$ (I always use the latter.)

free picture of a Normal density

They all have this shape:



density properties; calculations; applications

Maximum is attained at $x = \mu$ and inflection points are at $\mu \pm \sigma$ (basic calculus).

The density is symmetric around μ, which means $f(\mu - a) = f(\mu + a)$.

The Normal densities do not have closed-form anti-derivates (as usual) so a numerical routine is used (as usual).

You'll also have to get used to using tables.

Direct use: modeling things that are the result of "sums" of small independent contributions.

But the indirect use is far more important, and is the reason why they get all the love rather than the Poisson process family. (More later.)

N(0,1) calculation examples

Suppose $Z \sim N(0,1)$. ("The standard normal"). The book includes a table of $P(Z \leq z_p)$ for $z_p \in \{0.00, 0.01, \dots, 3.49\}$ with $P(Z \leq z_p) = p$.

From this table, for example:

$$P(Z \leq 1) = 0.8413 = P(Z \geq -1)$$

$$P(-1 < Z < 1) = 0.6827$$

$$P(Z \leq z_p) = 0.91 \iff z_p \approx 1.34$$

and plenty more. Become handy with this table.

"direct" normal model example

The SAT is a standardized test taken mainly by USA high school students, and some others.

The scores can be modeled using a normal distribution (why?) with, according to the May 2016 report, mean 1020 and standard deviation 194.

What is the probability that a student will have a score that exceeds 1250?

The answer is 0.1179 according to my computer.

But we don't have a N(1020, 194) table, which leads us to…

distributions of functions of random variables (with a focus on continuous r.v.)

functions of random variables: why?

Some of your calculus life was spent pondering things like $f(g(x))$ and their derivatives and integrals.

Now we'll ponder things like $g(X)$ focusing, of course, on the resulting distributions.

A common simple example would be a linear transformation ("unit change") with $g(x) = a + bx$ and:

$$Y = g(X) = a + bX$$

Given the distribution of X, what will be the distribution of Y?

"DOFORV" - "distribution of a function of a random variable"

DOFORV method I - use the cdf

The cdf is one way to characterize a distribution.

For example, suppose we have the cdf of X given by $F_x(x) = P(X \leq x)$ and we want the distribution of $Y = g(X) = a + bX$ for $b \neq 0$.

If we can find the cdf of Y, that gives us what we want:

$$F_Y(y) = \begin{cases} F_X\left(\frac{y-a}{b}\right) & : \text{ if } b > 0 \\ 1 - F_X\left(\frac{y-a}{b}\right) & : \text{ if } b < 0 \end{cases}$$

Then differentiate for density:

$$f_Y(y) = f_X\left(\frac{y-a}{b}\right) \cdot \frac{1}{|b|}$$

example: the normal distributions

Suppose $X \sim N(\mu, \sigma^2)$ and consider $g(x) = \frac{x-\mu}{\sigma}$. Determine the distribution of $Z = g(X)$.

We can go straight to the density to see $Z \sim N(0, 1)$.

Now we can use the table to finish the SAT example, in which $X \sim N(1020, 194)$ and we wanted $P(X > 1250)$.

$Z \sim N(0, 1)$ is called the "standard normal distribution". Its cdf gets its own special symbol: $F_z(z) = P(Z \leq z) = \Phi(z)$.

Apparently its density gets to be called $\phi(z)$ but this isn't as common.

Exercise: if Z is standard normal, find the distribution of $X = \mu + \sigma Z$.

example: uniform

Suppose $X \sim \mathrm{Unif}[0,1]$. Determine the distribution of $Y = a + bX$ with $b \neq 0$.

Exercise: Suppose $X \sim \mathrm{Unif}[a,b]$. Find c and d so that $Y = c + dX$ is $\mathrm{Unif}[0,1]$.

Exercise: Suppose $X \sim \mathrm{Exp}(1)$. Determine the distribution of $Y_1 = \lambda X$ and $Y_2 = a + \lambda X$. Draw pictures of them. Do consider $\lambda < 0$ (perhaps without any obvious application but still valid mathematically.)

(Note: the last part will result in a distribution we haven't given a name to, which is fine.)

DOFORV method I - cdf of square of r.v.

Squaring a random variable will turn out to have important applications.

Suppose X has cdf F_X and density f_X, and $g(x) = x^2$. Determine the distribution of $Y = g(X) = X^2$ using the cdf method.

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right)$$

Example: $Z \sim N(0,1)$ and $Y = Z^2$. Then Y has a $\text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution, a.k.a. χ_1^2, with a bonus fun fact for free.

$Y = Z^2 \sim \chi_1^2$ derivation

$$\begin{aligned}
f_Y(y) &= \frac{1}{2\sqrt{y}}\left(f_Z(\sqrt{y}) + f_Z(-\sqrt{y})\right) \\
&= \frac{1}{\sqrt{y}} f_Z(\sqrt{y}) \quad \text{(not always true - used symmetry of normal density)} \\
&= \frac{1}{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\sqrt{y})^2} \\
&= \frac{\frac{1}{2}^{\frac{1}{2}}}{\sqrt{\pi}} y^{\frac{1}{2}-1} e^{-y/2}
\end{aligned}$$

Bonus free fun fact: $\Gamma(\alpha) = \sqrt{\pi}$.

DOFORV method 2 - direct "theorem"

I am delighted that the book downplays this method as not as easy to use. Nor do I recommend it for practical use.

Theorem: Given X with density $f_X(x)$ and g monotone and differentiable with inverse g^{-1} where $f_X(x) > 0$, let $Y = g(X)$. Then:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

This theorem can be extended to non-monotonic g.

Proof: …

DOFORV - two other proofs

The theorem can be proved in two other ways that are straight outta calculus.

The proofs indeed all look very similar.

Proof 2: uses the "change of variables" method from integration (emphasizes my advice to always think of a density as living in an integral.)

Proof 3: uses the fundamental theorem of calculus.

a seemingly strange example

The techniques apply to any continuous r.v. X and to any differentiable, invertible $g(x)$.

So let's consider $X \sim \text{Exp}(\lambda)$ and let $g(x) = 1 - e^{-\lambda x}$. It turns out $Y \sim \text{Unif}[0, 1]$.

The function g was not chosen by accident—it it precisely the cdf $F_x(x)$ of X.

Theorem: If X is continuous and has cdf $F_x(x)$ then $Y = F_x(X)$ will have a uniform distribution on $[0, 1]$.

Proof: …

DOFORV - distributions of functions of (continuous) random variables - continued

DOFORV method 2 - direct "theorem"

I am delighted that the book downplays this method as not as easy to use. Nor do I recommend it for practical use.

The cdf approach is usually the cleanest and least error prone.

Theorem: Given X with density $f_X(x)$ and g monotone and differentiable with inverse g^{-1} where $f_X(x) > 0$, let $Y = g(X)$. Then:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

This theorem can be extended to non-monotonic g.

DOFORV - three proofs

The theorem can be proved using the cdf approach and two other ways that are straight outta calculus.

The proofs indeed all look very similar.

Proof 2: uses the "change of variables" method from integration (emphasizes my advice to always think of a density as living in an integral.)

Proof 3: uses the fundamental theorem of calculus (note: I never bothered with this one.)

added after class - Proof 2 - change of variables - I

In this proof I emphasize the point that the density lives in a definite integral and that the distribution of $Y = g(X)$ is obtained via a change of variables $y = g(x)$ inside the definite integral.

For $a < b$ the distribution of X with density f_X is characterized by:

$$\int_a^b f_X(x)\, dx,$$

which after the change of variable $y = g(x)$ becomes:

$$\int_{g^{-1}(a)}^{g^{-1}(b)} f_X(g^{-1}(y))\, \frac{d}{dy} g^{-1}(y)\, dy$$

added after class - Proof 2 - change of variables - II

Note that when g is decreasing, so will g^{-1}, which will mean $g^{-1}(a) > g^{-1}(b)$ and in that case the integral would be negative. So:

$$\int_a^b f_x(x)\,dx = \begin{cases} \int_{g^{-1}(a)}^{g^{-1}(b)} f_x(g^{-1}(y))\,\frac{d}{dy}g^{-1}(y)\,dy & : g \text{ increasing} \\ -\int_{g^{-1}(a)}^{g^{-1}(b)} f_x(g^{-1}(y))\,\frac{d}{dy}g^{-1}(y)\,dy & : g \text{ decreasing} \end{cases}$$

$$= \int_{g^{-1}(a)}^{g^{-1}(b)} f_x(g^{-1}(y))\,\left|\frac{d}{dy}g^{-1}(y)\right|\,dy$$

The integrand is the density for Y.

a seemingly strange example

The techniques apply to any continuous r.v. X and to any differentiable, invertible $g(x)$.

So let's consider $X \sim \mathrm{Exp}(\lambda)$ and let $g(x) = 1 - e^{-\lambda x}$. It turns out $Y \sim \mathrm{Unif}[0, 1]$.

The function g was not chosen by accident—it it precisely the cdf $F_x(x)$ of X.

Theorem: If X is continuous and has cdf $F_x(x)$ then $Y = F_x(X)$ will have a uniform distribution on $[0, 1]$.

Proof: $\ldots$

another DOFORV example

Suppose $X \sim \text{Gamma}(\alpha, \lambda)$ and

$$g(x) = \begin{cases} \frac{1}{x} & : x > 0, \\ 0 & : \text{otherwise.} \end{cases}$$

Determine the distribution of $Y = g(X)$ by finding its density.

probabilities involving more than one random variable at a time

motivation

A random variable is a function of a sample space, and we care about its distribution.

So far we've focussed on $X : S \to \mathbb{R}$.

Now we will look at $X : S \to \mathbb{R}^n$, which arise quite naturally. We've actually been doing this already sometimes, without saying so explicitly (see Case 2).

Case 1: you actually observe multiple things about a particular random outcome (e.g. you measure the weight and blood pressure of a randomly selected study participant.)

Case 2: you are considering a sequence of random variables that "replicate" the same "experiment" (e.g. repeat a Bernoulli(p) trial n times…)

discrete motivating example

Toss two fair six-sided dice. The sample space has 36 elements.

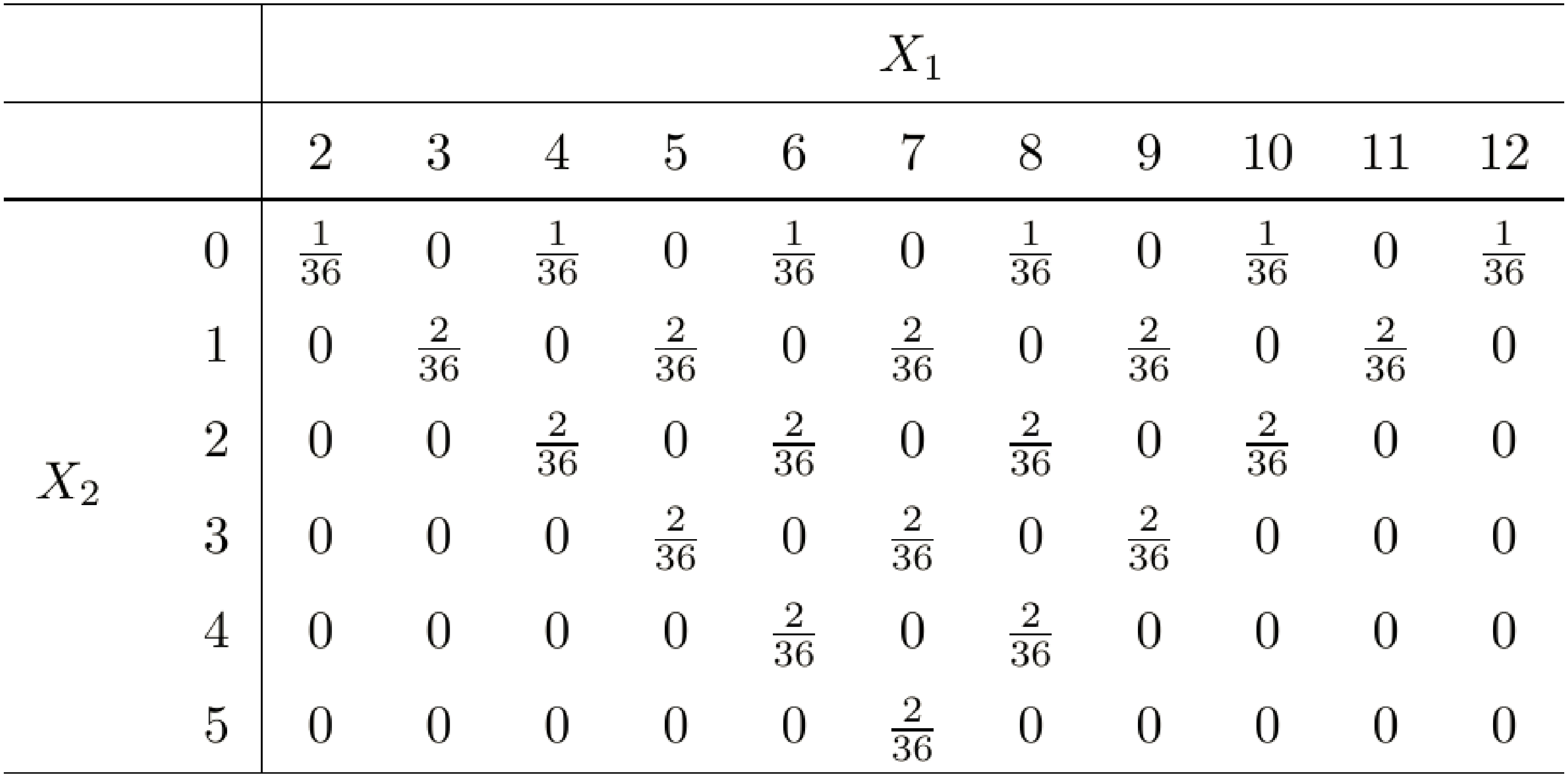
Observe the total and the difference.

Denote the total by X_1 and the difference by X_2, and $X = (X_1, X_2)$

$X : S \to \mathbb{R}^2$ is a random variable and we can consider probabilities of the form $P(X \in A_1 \times A_2) = P(X_1 \in A_1, X_2 \in A_2)$ for $A_i \subset \mathbb{R}$. All such probabilities together form the "distribution" of X.

But this is an excess of formality. Normally we work directly with the components, in this case X_1 and X_2, which are two discrete rvs and we can put probabilities of combinations of outcomes in a table.

table of probabilities



		X_1										
		2	3	4	5	6	7	8	9	10	11	12
X_2	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$
	1	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0
	2	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0
	3	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0
	4	0	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	0
	5	0	0	0	0	0	$\frac{2}{36}$	0	0	0	0	0

"joint" distribution

<Insert cannabis joke here>

Such a table summarizes the distribution of X, which we'll just call the joint distribution of X_1 and X_2.

The table has all the values of the form $p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$ and is called the joint probability mass function.

A joint pmf is non-negative, and its positive values sum to 1 (just like before).

The joint cdf is defined as: $P(X_1 \leq x_1, X_2 \leq x_2)$ for all $(x_1, x_2) \in \mathbb{R}^2$.

It is non-decreasing and right-continuous in both variables, and goes to 0 and 1 as both dimensions go to $\pm\infty$ respectively.

marginal distributions

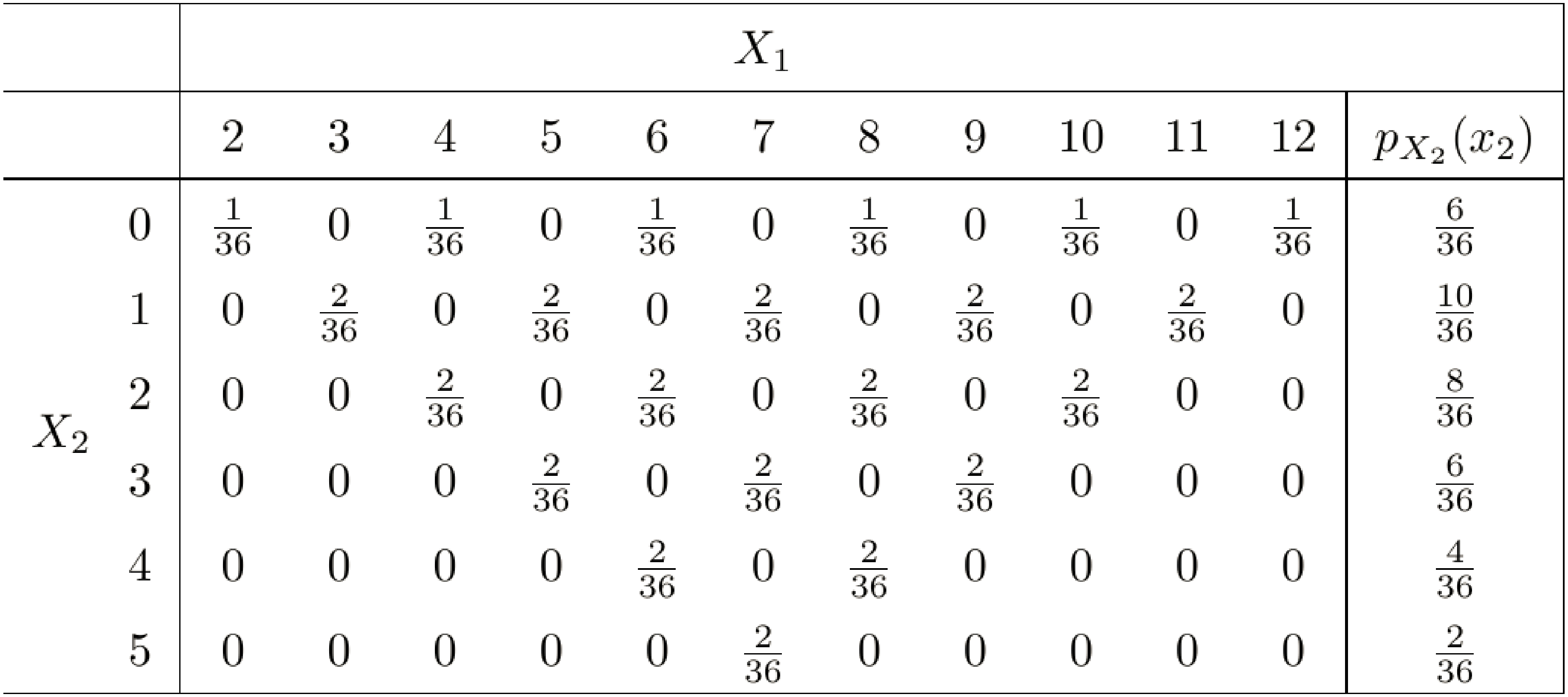
The pmf of any of the component random variables can be recovered by summing over all the others, e.g.:

$$p_{X_1}(x_1) = \sum_{x_2} p(x_1, x_2),$$

(where $\sum_{x_2}$ denotes summing over all values that X_2 takes on.)

This is called a marginal pmf, which characterizes the distribution of that random variable.

example: marginal for X_2



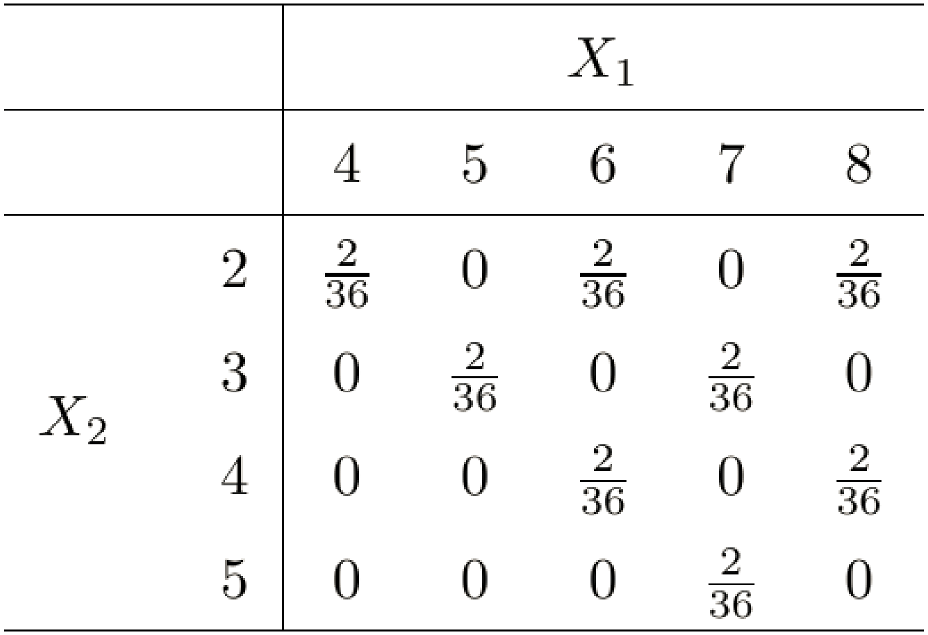
		X_1											
		2	3	4	5	6	7	8	9	10	11	12	$p_{X_2}(x_2)$
X_2	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	$\frac{6}{36}$
	1	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{10}{36}$
	2	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	$\frac{8}{36}$
	3	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	$\frac{6}{36}$
	4	0	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	0	$\frac{4}{36}$
	5	0	0	0	0	0	$\frac{2}{36}$	0	0	0	0	0	$\frac{2}{36}$

volumes under surfaces

discrete analogue

First, consider calculating in the discrete example:

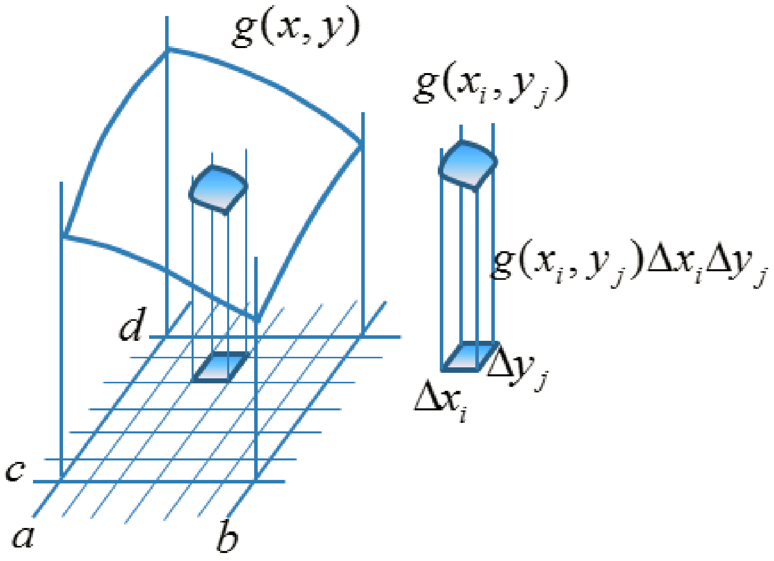
$$P(4 \leq X_1 \leq 8,\ 2 \leq X_2 \leq 5)$$



		X_1				
		4	5	6	7	8
X_2	2	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$
	3	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0
	4	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$
	5	0	0	0	$\frac{2}{36}$	0

"double integral" crash course - I

A double ("Riemann") integral calculates the volume between the a rectangle $[a,b]\times[c,d]$ in the xy–plane and a function $g(x,y)$ in pretty much the same way $\int g(x)$ calculates the area under a curve.

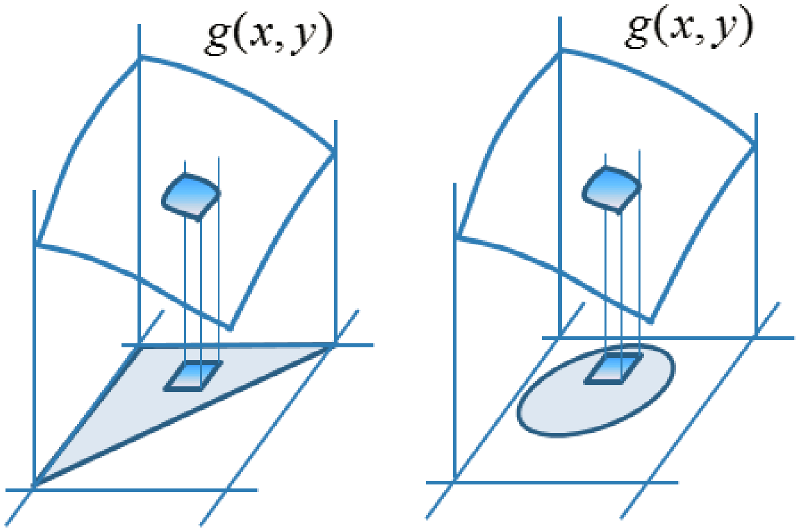


$$\int_a^b \int_c^d g(x,y)\,dx\,dy \approx$$

$$\sum_{i,j} g(x_i,y_i)\Delta x_i \Delta y_i$$

double integral crash course - II

In general the same idea applies to a volume between any region $B \subset \mathbb{R}^2$ and $g(x, y)$.



$$\iint_B g(x, y)\, dx\, dy$$

double integrals crash course - III

Two basic properties:

$$\iint\limits_{B} \left[\alpha_1 g_1(x,y) + \alpha_2 g_2(x,y)\right]\, dx\, dy =$$

$$\alpha_1 \iint\limits_{B} g_1(x,y)\, dx\, dy + \alpha_2 \iint\limits_{B} g_2(x,y)\, dx\, dy$$

When B_1 and B_2 are disjoint:

$$\iint\limits_{B_1 \cup B_2} g(x,y)\, dx\, dy =$$

$$\iint\limits_{B_1} g(x,y)\, dx\, dy + \iint\limits_{B_2} g(x,y)\, dx\, dy$$

double integrals crash course - IV

The actual calculation is just done one variable at a time, from the inside out (first holding the outside variable "constant")

Simple example (B is a rectangle): $g(x, y) = xy^2$. Consider:

$$\int_0^2 \int_1^3 xy^2 \; dx \, dy$$

And it makes no difference which order, but be careful when B is not a rectangle. Consider the integral of g over the region bounded by $x = 0$, $y = 0$, and $x + y = 1$.

joint continuous distributions

Reconsider measuring the weight X and (systolic) blood pressure Y of a randomly selected study participant. We'll be interested in things like:

$$P(70 < X < 80, 120 < Y < 140)$$

In this case X and Y are both continuous random variables. A fancy way to re-write the above would be to let $A = [70, 80] \times [120, 140]$ and get:

$$P((X, Y) \in A)$$

Definition: X and Y are jointly continuous if there is a $f(x, y)$ such that, for "all" $A \subset \mathbb{R}^2$

$$P((X, Y) \in A) = \iint\limits_A f(x, y)\, dx\, dy$$

joint density

The function $f(x, y)$ is called the joint density, and is used to calculate probabilities.

Properties: $f \geq 0$ and $\iint\limits_{\mathbb{R}^2} f(x, y)\, dx\, dy = 1$

Example 1: $f(x, y) = 1$ on $0 < x,\ y < 1$ and 0 otherwise. ("Joint Uniform") Is this a density? Calculate: $P(X < 1/2, Y < 1/2)$ and $P(X < Y)$.

Example 2: (artificial) $f(x, y) = cxy$ on $0 < x < 1$, $0 < y < 2$, and 0 otherwise. Determine c. Calculate $P(X > 0.5, 0 < Y < 1)$ and $P(Y > X)$.

joint cdf - continuous case

The joint cdf $F(x, y) = P(X \leq x, Y \leq y)$ can be calculated:

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv$$

It is a mystery of multivariable calculus how to obtain f from F

"partial" derivatives crash course - I

Maybe you got this far in your co-requisite!

With a function $g(x, y)$ you can take the derivative with respect to one variable at a time, holding the other variable constant. Notation:

$$\frac{\partial}{\partial x} g(x, y) \quad \text{and} \quad \frac{\partial}{\partial y} g(x, y).$$

When g is "smooth" you get the nice result:

$$\frac{\partial}{\partial y}\left[\frac{\partial}{\partial x} g(x, y)\right] = \frac{\partial}{\partial x}\left[\frac{\partial}{\partial y} g(x, y)\right],$$

and we just call this:

$$\frac{\partial^2}{\partial x \partial y} g(x, y).$$

joint cdf to joint density

Just take all the "partial" derivatives in any order you like.

$$\frac{\partial^2}{\partial x \partial y} F(x, y) = f(x, y)$$

"Proof: …"

Examples can be challenging! Consider $f(x, y) = xy$ on $0 < x < 1$, $0 < y < 2$ (…to be revisited…)

marginal cdf and marginal density

Just like in the discrete case we can recover information about X and Y individually by "integrating out" the other variable. The marginal densities are:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

What about the marginal cdfs?

Continue the example on the previous slide.

picture of F(x,y)

joint/marginal pdf example

Example D from the book.

$$f(x,y) = \begin{cases} \lambda^2 \exp(-\lambda y) & : 0 \le x \le y,\ \lambda > 0 \\ 0 & : \text{ otherwise.} \end{cases}$$

Exercise: review Example E from the book "Bivariate Normal". We will revisit this example.

independent random variables (which the book gets a bit wrong)

recall some definitions and results

Events A and B are independent if $P(A \cap B) = P(A)P(B)$.

$A \perp B \iff A^c \perp B \iff A \perp B^c \iff A^c \perp B^c$

"Experiments" $\mathcal{E}_A = \{A_1, A_2, \dots\}$ and $\mathcal{E}_B = \{B_1, B_2, \dots\}$ are independent if $A_i \perp B_j$ for all i, j.

Definition: Random variables X and Y are independent if:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any $A, B \subset \mathbb{R}$.

the book's definition is actually a theorem

Theorem: $X \perp Y$ if and only if the joint cdf $F(x, y) = F_X(x)F_Y(y)$ is the product of the marginal cdfs.

Proof: $\Longleftarrow$ ("only if") too hard; $\Longrightarrow$ left as exercise.

Corollary: $X \perp Y$ if and only if the joint $f(x, y) = f_X(x)f_Y(y)$

To verify, in practice check two things:

1. The density factors.
2. The non-zero region is a rectangle (possibly infinite in either direction.)

examples

1. $f(x, y) = xy$ on $0 < x < 1$ and $0 < y < 2$.

2. $f(x, y) = \lambda^2 \exp(-\lambda y)$ on $0 < x < y < \infty$.

3. $f(x, y) = \lambda^2 y \exp(-\lambda(x + y))$ on $x > 0$ and $y > 0$. Was changed to λ^2 from λ^3 in class.

from last time

Theorem: $X \perp Y$ if and only if the joint cdf $F(x, y) = F_X(x)F_Y(y)$ is the product of the marginal cdfs.

Proof: $\Longleftarrow$ ("only if") too hard; $\Longrightarrow$ left as exercise.

Corollary: $X \perp Y$ if and only if the joint $f(x, y) = f_X(x)f_Y(y)$

To verify, in practice check two things:

1. The density factors. Note: enough to factor into a function of x and a function of y.
2. The non-zero region is a rectangle (possibly infinite in either direction.) Note: technically a "cross product" is all that is needed, but in almost all practical cases it will be a rectangle.

other important independence results (advanced)

Theorem: If X and Y are independent, so are $g(X)$ and $h(Y)$ for any functions g and h.

Sketch of proof: …

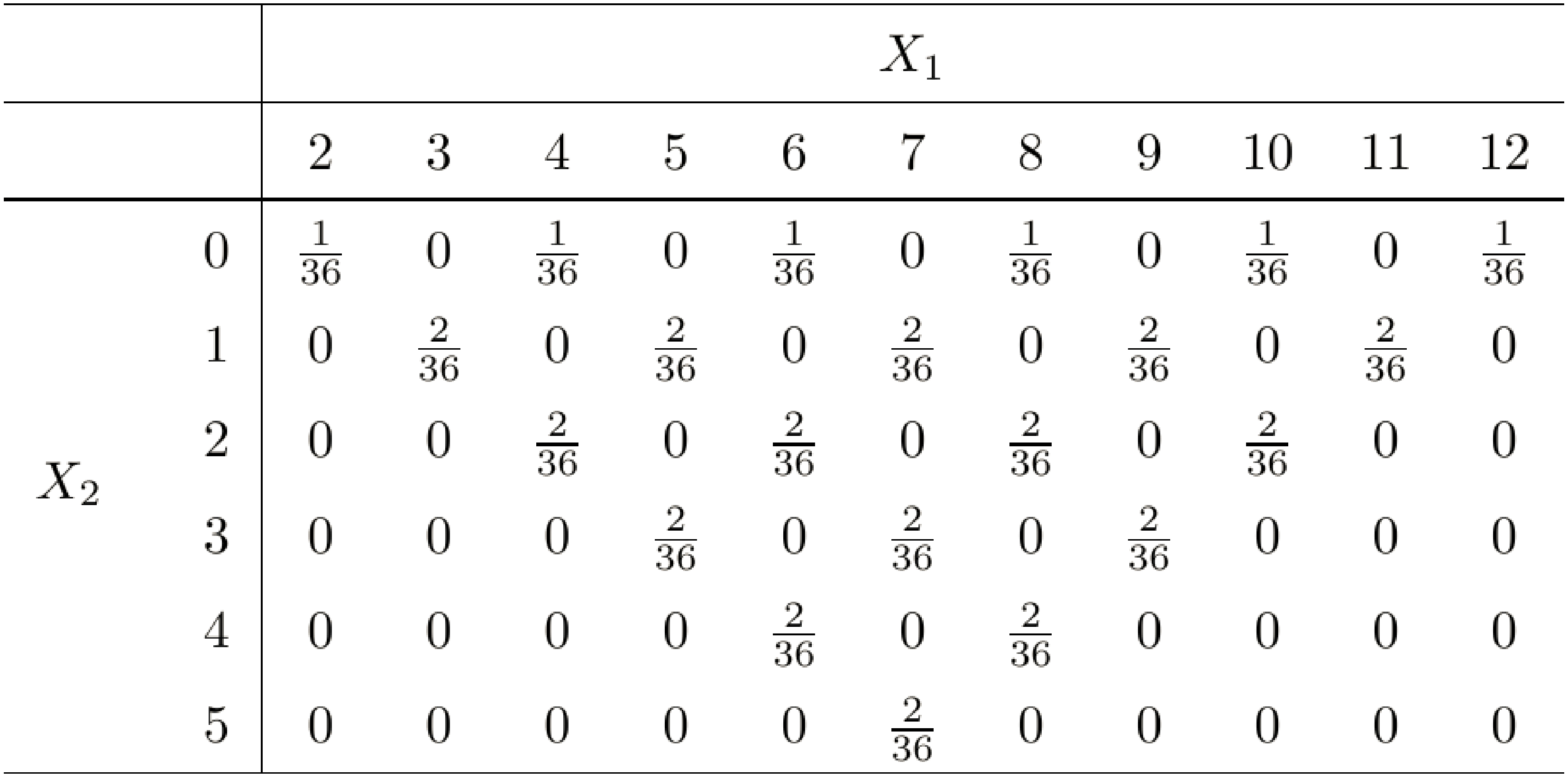
Definition of independence extends to any number of random variables. We say $X_1, X_2, \ldots, X_n$ are independent if:

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

for any subsets $A_i \in \mathbb{R}$.

conditional distributions

Recall the sum X and the absolute difference X of two dice:



		X_1										
		2	3	4	5	6	7	8	9	10	11	12
X_2	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$
	1	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0
	2	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0
	3	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0
	4	0	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	0
	5	0	0	0	0	0	$\frac{2}{36}$	0	0	0	0	0

discrete case

Given a joint pmf for X and Y denoted by $p(x, y)$, define:

$$p_{X|Y}(x|y) = \begin{cases} \frac{p(x,y)}{p_Y(y)} & : \text{ where } p_Y(y) > 0 \\ 0 & : \text{ otherwise} \end{cases}$$

For any fixed Y with $p_Y(y) > 0$, this is a valid pmf.

This pmf describes what is called "the conditional distribution of X given $Y = y$."

Useful result:

$$\begin{aligned} p(x, y) &= p_{X|Y}(x|y) p_Y(x) \\ p_X(x) &= \sum_y p_{X|Y}(x|y) p_Y(y) \end{aligned}$$

classic example

At home my phone rings Y times with $Y \sim \text{Poisson}(\lambda)$ in one hour.I answer the phone with probability p when it rings. What is the distribution of the X, the number of times I answer the phone in an hour?

(In fact p so $X = 0$ always. Note for people not in attendance…this is a joke about me not answering the phone.)

continuous case

The concept is similar. We examine a "slice" of the joint density at, say $X = x$ and consider the distribution of Y at that fixed value of x.

The conditional density of Y given $X = x$ is defined as:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

wherever $f_X(x) > 0$. Note: what is meant by this is $f_{X|Y}$ is only even defined whenever $f_X(x) > 0$. One still needs to carefully consider the support of $f_{Y|X}$.

Examples:

1. $f(x,y) = \frac{1}{\pi}$ on $x^2 + y^2 \leq 1$.
2. $f(x,y) = \lambda^2 e^{-\lambda y}$ on $0 < x < y$.

from last time

Theorem: $X \perp Y$ if and only if the joint cdf $F(x, y) = F_X(x)F_Y(y)$ is the product of the marginal cdfs.

Proof: $\Longleftarrow$ ("only if") too hard; $\Longrightarrow$ left as exercise.

Corollary: $X \perp Y$ if and only if the joint $f(x, y) = f_X(x)f_Y(y)$

To verify, in practice check two things:

1. The density factors. Note: enough to factor into a function of x and a function of y.
2. The non-zero region is a rectangle (possibly infinite in either direction.) Note: technically a "cross product" is all that is needed, but in almost all practical cases it will be a rectangle.

other important independence results (advanced)

Theorem: If X and Y are independent, so are $g(X)$ and $h(Y)$ for any functions g and h.

Sketch of proof: …

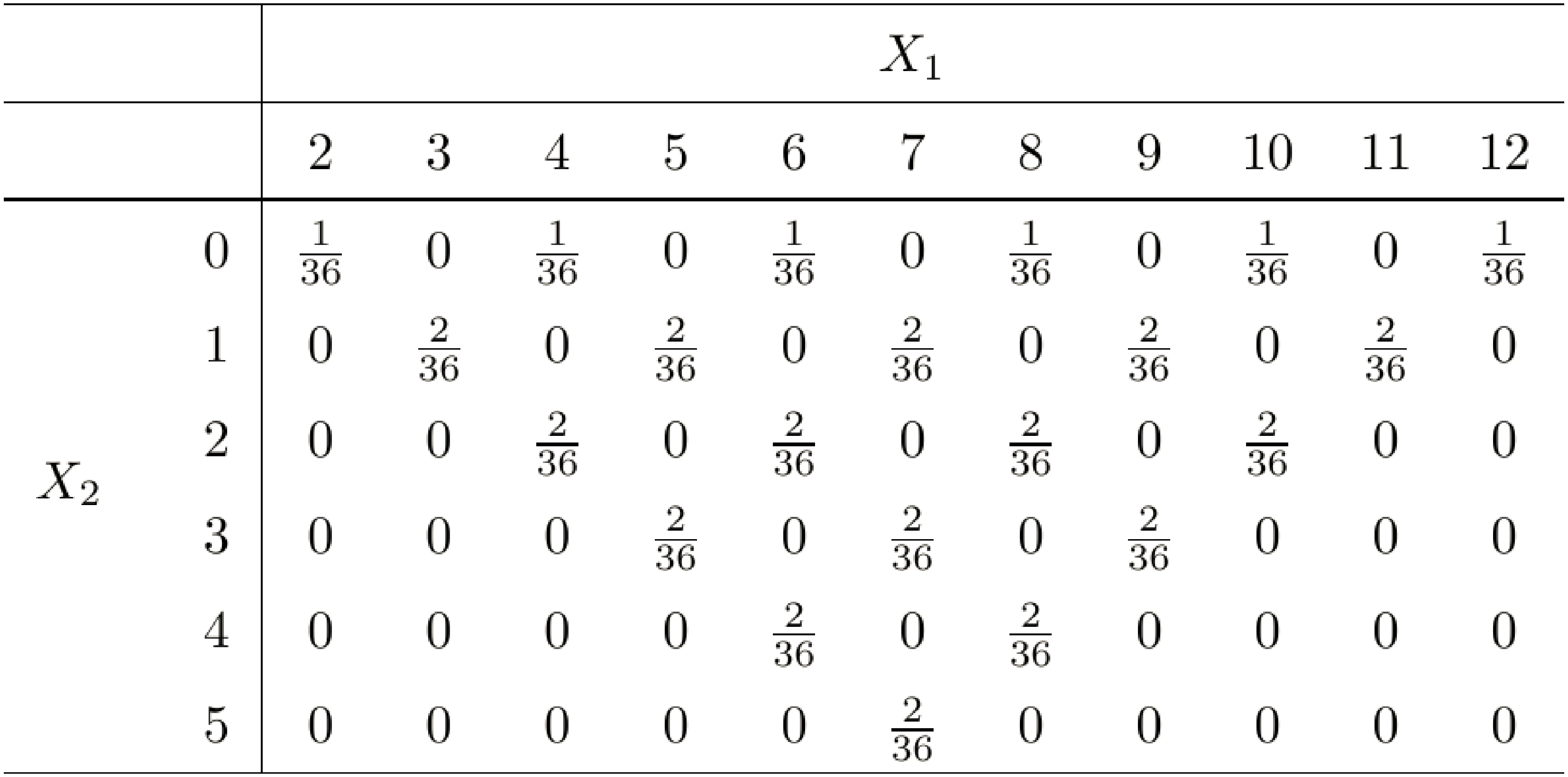
Definition of independence extends to any number of random variables. We say $X_1, X_2, \dots, X_n$ are independent if:

$$P(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

for any subsets $A_i \in \mathbb{R}$.

conditional distributions

Recall the sum X and the absolute difference X of two dice:



		X_1										
		2	3	4	5	6	7	8	9	10	11	12
X_2	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$
	1	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0
	2	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0
	3	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0
	4	0	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	0
	5	0	0	0	0	0	$\frac{2}{36}$	0	0	0	0	0

discrete case

Given a joint pmf for X and Y denoted by $p(x, y)$, define:

$$p_{X|Y}(x|y) = \begin{cases} \frac{p(x,y)}{p_Y(y)} & : \text{ where } p_Y(y) > 0 \\ 0 & : \text{ otherwise} \end{cases}$$

For any fixed Y with $p_Y(y) > 0$, this is a valid pmf.

This pmf describes what is called "the conditional distribution of X given $Y = y$."

Useful result:

$$\begin{aligned} p(x, y) &= p_{X|Y}(x|y)p_Y(x) \\ p_X(x) &= \sum_y p_{X|Y}(x|y)p_Y(y) \end{aligned}$$

classic example

At home my phone rings Y times with $Y \sim \text{Poisson}(\lambda)$ in one hour.I answer the phone with probability p when it rings. What is the distribution of the X, the number of times I answer the phone in an hour?

(In fact p so $X = 0$ always. Note for people not in attendance…this is a joke about me not answering the phone.)

continuous case

The concept is similar. We examine a "slice" of the joint density at, say $X = x$ and consider the distribution of Y at that fixed value of x.

The conditional density of Y given $X = x$ is defined as:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

wherever $f_X(x) > 0$. Note: what is meant by this is $f_{X|Y}$ is only defined whenever {f_{\tiny{_}}}X(x)=0. One still needs to carefully consider the support of $f_{Y|X}$..

Examples:

1. $f(x,y) = \frac{1}{\pi}$ on $x^2 + y^2 \leq 1$.
2. $f(x,y) = \lambda^2 e^{-\lambda y}$ on $0 < x < y$.

the bivariate normal distributions - an important class of joint distributions

since civilization is over anyway…

Let's do something crazy. Recall $X \sim N(\mu, \sigma^2)$ has density for all $x \in \mathbb{R}$:

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

Re-imagine x and μ as a column vectors with one element each: $\mathbf{x} = \left(x \right)$ and $\mu = \left(\mu \right)$. Re-imagine σ^2 as a 1×1 matrix $\mathbf{\Sigma} = \left(\sigma^2 \right)$.

Note that $\det \mathbf{\Sigma} = |\mathbf{\Sigma}| = \sigma^2$ and $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}$, and:

$$f_X(x) = \frac{1}{|\mathbf{\Sigma}|^{1/2} (2\pi)^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \right)$$

bivariate normal

The random variables X_1 and X_2 have a bivariate normal distribution with paramaters μ_1, μ_2, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, and $-1 < \rho < 1$ if:

$$f(x_1, x_2) = \frac{1}{|\mathbf{\Sigma}|^{1/2}(2\pi)^{2/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^t \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \right)$$

where:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

book version of density

Some work will reveal that this is equivalent to the formula given in the textbook.

These densities actually look like "bells".

What if $\rho = 0$?

The densities have the interesting properties that the marginal distributions are normal, and the conditional distributions are also normal.

the classic density formulae: $W_1 = X + Y$

$$\begin{aligned} f_{W_1}(w) &= \int_{-\infty}^{\infty} f(x, w-x)\,dx \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\,dx \qquad (\text{when } X \perp Y) \end{aligned}$$

Proof:…

Example: $X \sim N(0,1)$ and $Y \sim N(0,1)$ with $X \perp Y$.

Example: $X \sim \mathrm{Unif}[0,1]$ and $Y \sim \mathrm{Unif}[0,1]$ with $X \perp Y$.

Exercise (textbook example): $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\lambda)$ with $X \perp Y$

the classic density formulae: $W_2 = Y \,/\, X$

$$\begin{aligned} f_{W_2}(w) &= \int_{-\infty}^{\infty} f(x, wx)|x|\, dx \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(wx)|x|\, dx \qquad (\text{when } X \perp Y) \end{aligned}$$

NOTE ON PROOF IN CLASS: My notes were correct. Something did not translate to the board. It turns out I used exactly the same method of proof as the textbook. So you can look on page 98 and everything is there.

"Mandatory" exercise (done in book as example): $X \sim N(0,1)$ and $Y \sim N(0,1)$ with $X \perp Y$. This is a classic. We say W_2 has a Cauchy distribution with density $\frac{1}{\pi}\frac{1}{1+w^2}$

Example: $X \sim \mathrm{Exp}(\lambda)$ and $Y \sim \mathrm{Exp}(\lambda)$ with $X \perp Y$.

the minumum and maximum of n r.v.s

Practical example: A mine has 20 haul trucks with engines whose time-to-failure (from any cause) is random and with Exp(λ) distribution. How long until the first failure of any truck?

Assumption: engines fail independently.

Notation: $X_1, X_2, X_3, \ldots, X_{20}$ are independent (assumed) and have the same distribution Exp(λ).

"i.i.d.": independent and identically distributed

Nothing special about 20, so let's consider the case of $X_1, \ldots, X_n$ i.i.d. Exp(λ). The joint density will be:

$$f(x_1, \ldots, x_n) = \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n} = \lambda^n e^{-n\lambda \sum x_i}$$

minimum of n i.i.d. exponential r.v.s

Denote by $X_{(1)}$ the time to the first failure.

Theorem: $X_{(1)} \sim \mathrm{Exp}(n\lambda)$.

Proof: $\dots$

Exercise (general case): If $X_1, \dots, X_n$ are i.i.d. each with density $f(x)$ and cdf $F(x)$, the density of $X_{(1)} = \min\limits_{1 \le i \le n} \{X_1, \dots, X_n\}$ is:

$$f_{X_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1}$$

See section 3.7 up to the end of "E X A M P L E B"

Maximum is similar.

expected value

big money!

You and I agree to gamble on the outcome of one toss of one coin.

If H appears, I give you \$100. If T appears, you give me \$100.

This is a fair game.

Denote by Y my financial outcome. X is discrete with pmf:

$$P(Y = y) = \begin{cases} 0.5 & : y = 100 \\ 0.5 & : y = -100 \end{cases}$$

It's a fair game, so my "average" outcome should be 0. Otherwise it would not be rational for either of us to play the game!

This average is exactly $(100)(0.5) + (-100)(0.5) = 0.$

expected value - discrete case

Definition: The expected value of X that takes on values $\{x_1, x_2, \ldots\}$ with pmf $p(x)$ is:

$$E(X) = \sum_{i=1}^{\infty} x_i p(x_i)$$

provided $\sum_{i=1}^{\infty} |x_i| p(x_i) < \infty$. Otherwise $E(X)$ is undefined.

some of the "named" distributions

$X \sim \text{Bernoulli}(p)$. Then $E(X) = p$.

Proof: …

$X \sim \text{Binomial}(n, p)$. Then $E(X) = np$.

Proof: …

$X \sim \text{Geometric}(p)$. Then $E(X) = \frac{1}{p}$.

Proof: … (Use: $\frac{d}{dr} \sum_{x=0}^{\infty} r^x = \sum_{x=1}^{\infty} x r^{x-1}$ when $|r| < 1$.)

Exercise (book example): $X \sim \text{Poisson}(\lambda)$. Then $E(X) = \lambda$.

fun with expecations

Suppose X has pmf $p(x) = \frac{6}{\pi^2 x^2}$ on $x \in \{1, 2, 3, \ldots\}$.

Suppose X has pmf $p(x) = \frac{3}{\pi^2 x^2}$ on $x \in \{\pm 1, \pm 2, \pm 3, \ldots\}$.

Treat X as a constant, i.e. $X = a$. Then $E(X) = E(a) = a$.

For a sample space S and event $A \subset S$, consider the "indicator" random variable I_A. Then $E(I_A) = P(A)$.

expected value - continuous case

If X is continuous with density f, its expected value is:

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

provided $\int_{-\infty}^{\infty} |x|f(x)\,dx < \infty$.

Examples: $U \sim \mathrm{Unif}[a,b]$, $Z \sim N(0,1)$, X Cauchy,…

For a proof that the expected value (if it exists) of a random variable X with density symmetric around a must be $E(X) = a$, see Theorem 6.1.1 in the supplemental text: Schay (2016) Introduction to Probability with Statistical Applications

Exercise (textbook example): $X \sim \mathrm{Gamma}(\alpha, \lambda)$

optional: "meaning" of continuous definition

Not done in class. Take it or leave it.

The meaning of the discrete definition makes intuitive sense - $E(X)$ is the probability-weighted average of the possible X outcomes.

The continuous definition can be motivated by the notion of "discretizing" the range of X into a partition $\{x_1, x_2, \ldots\}$. and considering the following approximation, which is again a probability-weighted average:

$$E(X) \approx \sum x_i P(x_i < X \leq x_{i+1})$$

Draw a diagram of this to see what is going on. If the partition is fine enough we have $P(x_i < X \leq x_{i+1}) \approx f(x_i)(x_{i+1} - x_i) \approx f(x_i)\Delta x_i$. Pass to the limit:

$$E(X) \approx \sum x_i P(x_i < X \leq x_{i+1}) \approx \sum x_i f(x_i)\Delta x_i \longrightarrow \int x f(x)\, dx$$

$E(g(X))$ and extensions

Motivation: suppose $X \sim N(\mu, \sigma^2)$. What is $E(X)$? The answer is μ. Lots of ways to figure this out.

Using the density is tedious but do-able. Or we could use the fact that $X = \mu + \sigma Z$ with $Z \sim N(0, 1)$.

Theorem: Given X and $E(X)$ exists, consider $g(x) = a + bx$. Then $E(g(X)) = E(a + bX) = a + bE(X)$.

Proof: …

$E(g(X))$

A theorem which is too difficult to prove generally is: given X, any g, and $Y = g(X)$, then:

$$E(Y) = E(g(X)) = \begin{cases} \sum g(x)p(x) & : X \text{ discrete} \\ \int g(x)f(x)\,dx & : X \text{ continuous} \end{cases}$$

in both cases provided the sum/integral congerges "absolutely" (i.e. with $|g(x)|$.)

Example: Average volume of sphere with radius $R \sim \text{Exp}(1)\ldots$

$E(g(X_1, \dots, X_n))$

Some typical applications:

$$E(X_1 \cdot X_2)$$

$$E\left(X_1 + X_2\right)$$

$$E(X_1 + \dots + X_n)$$

$$E\left(\overline{X}\right) = E\left(\frac{X_1 + \dots + X_n}{n}\right)$$

Theorem (continuous version): $X_1, \dots, X_n$ have joint density $f(x_1, \dots, x_n)$ and $Y = g(X_1, \dots, X_n)$. Then:

$$E(Y) = \int \dots \int g(x_1, \dots, x_n) f(x_1, \dots, x_n) \, dx_1 \, \dots \, dx_n$$

examples

Suppose $X_1 \perp X_2$. Consider $E(X_1 \cdot X_2)\ldots$

Exercise: $X_1 \perp X_2$. Consider $E(g(X_1)\, h(X_2))$

Now suppose $X_1, \ldots, X_n$ are i.i.d. with $E(X_i) = \mu$. Consider:

$$E\left(\overline{X}\right) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu$$

Note added after class: Try the proof with $n = 2$. Notice that there was nothing special about n. Also notice that the only requirement for $E\left(\overline{X}\right) = \mu$ is just that $E(X_i) = \mu$. Try for $n = 2$ just with a general joint distribution and no independence and not the same distribution for X_1 and X_2 and only $E(X_1) = E(X_2) = \mu$

$X \sim \text{NegBin}(r, p)\ldots$

key point from end of last class, restated

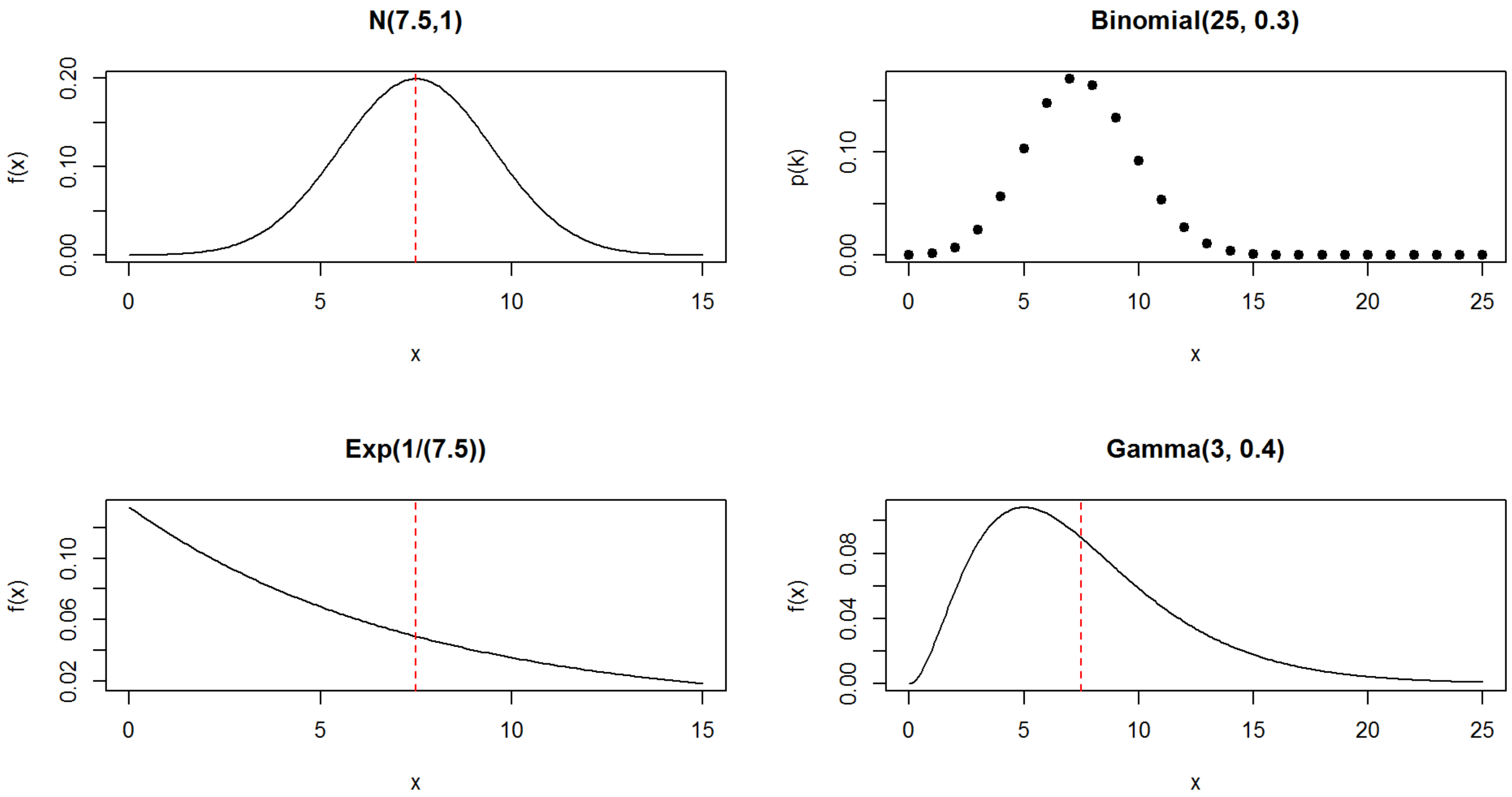
I did: $X_1, \dots, X_n$ i.i.d. implies $E\left(\overline{X}\right) = \mu$.

A different (better?) approach could have started with a more fundamental:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

which is true no matter what. (No independence required and all expected values can be different.)

looking back: getting a sense of $E(X) = 7.5$



N(7.5,1)

Binomial(25, 0.3)

Exp(1/(7.5))

Gamma(3, 0.4)

putting a number on variation

Expected value is a measure of "location", but random variables with the same "location" can be quite different.

Consider the coin tossing game with $E(Y) = 0$:

$$P(Y = y) = \begin{cases} 0.5 & : y = 100 \\ 0.5 & : y = -100 \end{cases}$$

One thing leads to another. Family trees are compared and contrasted, and after more than a few schnapps things get interesting:

$$P(Y_2 = y) = \begin{cases} 0.5 & : y = 1000 \\ 0.5 & : y = -1000 \end{cases}$$

Still, $E(Y_2) = 0$. But the values of Y_2 are more spread out.

variance

One way to measure spread is to use the variance of X, defined as:
$\mathrm{Var}(X) = E\left[(X - E(X))^2\right]$.

This is a use of $E(g(X))$ with $g(x) = (x - E(X))^2$.

Very useful, and almost always the way to perform the actual calculation.

$$\begin{aligned}
\mathrm{Var}(X) &= E\left(X^2 - 2XE(X) + E(X)^2\right) \\
&= E\left(X^2\right) - 2E(X)E(X) + E(X)^2 \\
&= E\left(X^2\right) - E(X)^2.
\end{aligned}$$

Note: existence of $\mathrm{Var}(X)$ requires existence of both $E(X^2)$ and $E(X)$.

Fun fact: existence of $E(X^2)$ implies the existence of $E(X)$.

examples, sketches, exercises, hints

$X \sim \text{Bernoulli}(p)$ ($p(1-p)$)

$Z \sim N(0,1)$ (1)

$X \sim \text{Poisson}(\lambda)$ (λ) (uses a trick!)

Variance of $X = a$ constant.

Basic examples for exercise (answer): Exp(λ) ($1/\lambda^2$), Gamma (α/λ^2), Geometric ($(1-p)/p^2$) (trick: differentiate power series twice), Binomial ($np(1-p)$)(use Poisson trick).

$\mathrm{Var}(a + bX)$, $\mathrm{Var}(X + Y)$ (independent case)

$\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$. Proof…

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

When $X \perp Y$, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. Proof…

Actually independence is stronger than necessary. Only needed $E(XY) = E(X)E(Y)$; to be revisited.

variance of the "sample average"

This is a "grand" example of particular importance.

Suppose again $X_1, \dots, X_n$ is i.i.d. with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$.. We already know $E\left(\overline{X}\right) = \mu$.

What about $\mathrm{Var}\left(\overline{X}\right)$?

standard deviation; notational conventions

Variance is in the square of the unit of measure of X.

"Standard deviation" is just:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)},$$

and is a more practical number to use for descriptive purposes (but less practical for theoretical developments.)

A common abbreviation for $\mathrm{Var}(X)$ is σ^2 so that $\mathrm{SD}(X) = \sigma$.

the Russians are coming!

$E(X)$ and $E\left(X^2\right)$ provide information about X that limit its values and probabilities to some extent. Two examples are Markov's and Chebyshev's inequalities.

Theorem (Markov): If $X \geq 0$ has expected value $E(X)$, then:

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Proof: Much easier than in book…(!)

Theorem (Chebyshev): If $\mathrm{Var}(X) = \sigma^2$ and $E(X) = \mu$:

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Proof: Apply Markov to the random variable $(X - \mu)^2$ and t^2.

Markov and Chebyshev examples

Consider $X \sim \mathrm{Exp}(5)$ and $t = 0.5$.

$E(X) = 1/5$ and $\mathrm{Var}(X) = 1/25$

$P(X \geq 0.5) = 0.082085 \leq 0.4$ (Markov)

$P(|X - 1/5| \geq 0.5) = 0.0301974 \leq 0.16$ (Chebyshev)

Our Russian friends more useful in theory than in practice.

a quick tour of covariance, correlation, and conditional expectation

covariance, and correlation

The quantity we saw in $\mathrm{Var}(X+Y)$:

$$E(XY) - E(X)E(Y) = E((X - E(X))(Y - E(Y)))$$

is called covariance or $\mathrm{Cov}(X, Y)$, and is a measure of linear association between the distributions of X and Y.

Note: $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.

Note: Covariance is "multi-linear", which means linear in both variables.

meaning of "linear association" of X and Y

Essentially: when X exceeds $E(X)$, is Y likelier, or not, to exceed $E(Y)$?

Consider some examples with X and Y jointly uniform over these triangles:

Bounded by (0,0), (0,1), and (1,0). ($\text{Cov}(X, Y) = -3/108$)

Bounded by (0,0), (0,1), and (-1,0). ($\text{Cov}(X, Y) = 3/108$)

correlation

Covariance is in the multiple of the X and Y units. A "unitless" version of covariance is another measure of linear association called "correlation", defined as:

$$\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Often denoted by ρ and often called "correlation coefficient."

Example (textbook section 4.3 EXAMPLE F) the ρ in the definition of the bivariate normal density is the correlation between X and Y.

conditional expectation given $Y = y$

First, recall the definition of conditional density for X given $Y = y$:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

This is a valid density, and we can consider the expected value of the random variable $X|Y = y$ with this density:

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx.$$

Analogous definition for the discrete case.

Nothing at all special about this.

recall the two dice example

Considering the sum and absolute difference:

		X_1										
		2	3	4	5	6	7	8	9	10	11	12
X_2	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$
	1	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0
	2	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0
	3	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0
	4	0	0	0	0	$\frac{2}{36}$	0	$\frac{2}{36}$	0	0	0	0
	5	0	0	0	0	0	$\frac{2}{36}$	0	0	0	0	0

more on two dice

(Done on board in class.)

It is easy to calculate the expected value of $X_2|X_1 = x_1$ for each $x_1 \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ (the ‖ is there instead of | because of a formatting issue):

x_1	2	3	4	5	6	7	8	9	10	11	12
$E(X_2\|X_1 = x_1)$	0	1	4/3	2	12/5	3	12/5	2	4/3	1	0

You can associate each $E(X_2\|X_1 = x_1)$ with the probability $P(X_1 = x_1)$ itself to define a new random variable that has the following probability mass function:

Probability:	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
Outcome:	0	1	4/3	2	12/5	3	12/5	2	4/3	1	0

The random variable with this p.m.f has a (confusing!) name: the conditional expectation of X_2 given X_1 or in notation: $E(X_2|X_1)$.

Despite its confusing name, it's just a random variable.

a little more on two dice

(Not done explicitly in class.)

It is easy to use the pmf on the previous page to show that the expected value of this odd new random variable is 35/18.

You can also go back to page 8 of the spooky 2016-10-31 notes and find the marginal pmf for X_2 and use that to calculate $E(X_2)$. You will find that this is also 35/18.

This is not a coincidence.

more cond. exp. not done in class

Let's build up to this apparently mysterious $E(X|Y)$ thing.

Start with: Any random variable Y.

Then consider: Any other random variable X. The conditional distribution of X given $Y = y$ can be characterized by (say) the conditional density $f_{X|Y}(x|y)$ and the expected value of this conditional distribuion is just:

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx.$$

Define a function $g : \mathbb{R} \to \mathbb{R}$: by considering another random variable X and defining $g(y) = E(X|Y = y)$

Define a new random variable: $W = g(Y)$

Give W a confusing name, because why not: Call W the "conditional expectation of X given Y" or $E(X|Y)$. It is not an expected value (i.e. a constant); it is a random variable, with a distribution etc.

interesting property not done in class

Since $g(Y) = E(X|Y)$ is a random variable we might be interested in some of the properties of its distribution.

One interesting property is its expected value.

$$\begin{aligned} E(E(X|Y)) = E(g(Y)) &= \int_{-\infty}^{\infty} g(y) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx \right] f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x\, \frac{f(x,y)}{f_Y(y)}\, dx \right] f_Y(y)\, dy \end{aligned}$$

(continued…)

…from previous

$$
\begin{aligned}
E(E(X|Y)) &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x\,\frac{f(x,y)}{f_Y(y)}\,dx\right] f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,\frac{f(x,y)}{f_Y(y)} f_Y(y)\,dy\,dx \qquad \text{(change order of integration)} \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y)\,dy\,dx \\
&= \int_{-\infty}^{\infty} x f_X(x)\,dx \\
&= E(X)
\end{aligned}
$$

covariance, correlation, conditional expectation in context

(not done in class…)

These topics are not central to this course, so I have given them only the briefest of introductions.

They will re-appear possibly in STA261 but definitely in things like regression, experimental design, sampling theory, and other areas of applied statistics.

Covariance and correlation also have empirical analogues when analyzing actual datasets.

Conditional expectation also plays an important theoretical role in advanced probability theory and in things like stochastic processes.

a new way to completely characterize the distribution of a random variable

what is important about a random variable?

One of the main points of this course: we essentially care about the distribution of a random variable X, which maps events $X \in A$ to probabilities $P(X \in A)$.

So far we have the following ways (only) to characterize a random variable's distribution, depending on the circumstances:

1. Cumulative distribution function $F(x) = P(X \leq x)$
2. Survival functions $S(x) = P(X > x) = 1 - F(x)$
3. (Discrete only) Probability mass functions $p(x)$
4. (Continuous only) Probability density functions $f(x)$.

We use whichever one is most convenient for a given situation.

mean, variance, and "moments"

$E(X)$ gives some (but not all) information about a random variable.

$\text{Var}(X) = E\left(X^2\right) - E(X)^2$ gives some more (but still not all) information about a random variable. (e.g. a Binomial(20, 0.5) and a N(10, 5) distribution have the same means and variances, but are not (at all!) the same distributions.)

Definition: For integers $k \geq 0$, $E\left(X^k\right)$ (if it exists) is called the kth moment of a random variable.

Note: calculating all these moments is not the point at all. The concept itself is what is important.

But a few examples could be $X \sim \text{Bernoulli}(p)$ and $Z \sim N(0, 1)$.

moment sequence characterizes

If turns out (proof beyond this course, although I'll give the reason in a minute) that if all moments exist:

$$\left\{E(X), E\left(X^2\right), E\left(X^3\right), E\left(X^4\right), \ldots\right\}$$

then this sequences sometimes gives a characterization of the distribution of X.

(I'll tell you how to know when this "sometimes" is, soon.)

But carrying around infinite sequences is not convenient. We need a nice, convenient package for them.

"generating function"

It often happens in mathematics that some big concept can be characterized by an infinite sequence of numbers.

A common trick is to use those numbers as coefficients in some infinite series (usually constructed to be otherwise useful in some way as well.)

With the moment sequence we shall do exactly that:

$$
\begin{aligned}
&= 1 + E(X)\frac{t}{1!} + E\left(X^2\right)\frac{t^2}{2!} + E\left(X^3\right)\frac{t^3}{3!} + \cdots \\
&= \sum_{k=0}^{\infty} E\left(X^k\right)\frac{t^k}{k!}
\end{aligned}
$$

Why this particular infinite series? Because I said so.

Let's call this function $m(t)$, which is a generating function using all moments.

$m(t)$ is actually an expected value itself

A bit of manipulation could give:

$$m(t) = \sum_{k=0}^{\infty} E\left(X^k\right)\frac{t^k}{k!} = E\left(e^{tX}\right)$$

Of particular importance is the existence of that radius of convergence $|t| < R$ (although the actual radius $R > 0$ doesn't matter.)

Summary: if all moments exist and if $E\left(e^{tX}\right)$ is defined with a positive radius of convergence, then we have a neatly packaged object containing all the moments.

Theorem: under these conditions, $m(t)$ completely characterizes the distribution of X.

Proof: Too hard. Based on the uniqueness of something called a Laplace transform; needs complex analysis.

a generating function using moments

Definition: $m(t) = E\left(e^{tX}\right)$ is called the moment generating function for X, if it exists in an interval containing 0.

Examples: Bernoilli, Binomial, $N(0,1)$

You can extract moments from the mgf using derivatives:

$$\frac{d^k}{dt^k} m(t) = \frac{d^k}{dt^k} E\left(e^{tX}\right) = E\left(\frac{d^k}{dt^k} e^{tX}\right) = E\left(X^k e^{tX}\right)$$

and then setting $t = 0$.

Examples…

the more important use of mgf

Suppose X and Y are independent random variables with mgfs $m_X(t)$ and $m_Y(t)$. Then $W = X + Y$ has mgf:

$$\begin{aligned} m_W(t) = m_{X+Y}(t) &= E\left(e^{t(X+Y)}\right) \\ &= E\left(e^{tX} e^{tY}\right) \\ &= E\left(e^{tX}\right)E\left(e^{tY}\right) = m_X(t)m_Y(t) \end{aligned}$$

Corollary: If $X_1, \ldots, X_n$ are independent and $W = \sum X_i$, then:

$$m_W(t) = \prod_{i=1}^{n} m_{X_i}(t)$$

some sums of r.v.s

If $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), then $\sum X_i \sim \text{Binomial}(n, p)\ldots$

If $X_1, \ldots, X_n$ are i.i.d. Geometric(p), then $\sum X_i \sim \text{NegBin}(n, p)\ldots$

This is fundamentally a "lookup table" technique.

Others (exercises):

sum of n independent Exp(λ) is Gamma(n, λ) sum of n independent Poisson(λ) is Poisson$(n\lambda)$ sum of $X_i \sim \text{Binomial}(n_i, p)$ is Binomial$\left(\sum n_i, p\right)$

the normal distributions

First, suppose X has mgf $m_X(t)$ and $Y = a + bX$. What is $m_Y(t)$?

So what is the mfg of a general $N(\mu, \sigma^2)$?

Finally, if $X_1, \dots, X_n$ are independent with $X_i \sim N(\mu_i, \sigma_i^2)$?, what distribution is $X = \sum X_i$?

some sums of r.v.s

If $X_1, \dots, X_n$ are i.i.d. Bernoulli(p), then $\sum X_i \sim \text{Binomial}(n, p)\dots$

If $X_1, \dots, X_n$ are i.i.d. Geometric(p), then $\sum X_i \sim \text{NegBin}(n, p)\dots$

This is fundamentally a "lookup table" technique.

Others (exercises):

sum of n independent Exp(λ) is Gamma(n, λ) sum of n independent Poisson(λ) is Poisson$(n\lambda)$ sum of $X_i \sim \text{Binomial}(n_i, p)$ is Binomial$\left(\sum n_i, p\right)$ distribution of sum of $X_i \sim \text{Binomial}(n_i, p_i)$ (different p_i!) cannot be determined using mgf technique.

the normal distributions

First, suppose X has mgf $m_{X}(t)$ and $Y = a + bX$. What is $m_{Y}(t)$?

So what is the mfg of a general $N(\mu, \sigma^2)$?

Finally, if $X_1, \ldots, X_n$ are independent with $X_i \sim N(\mu_i, \sigma_i^2)$?, what distribution is $X = \sum X_i$?

sequences of random variables, convergence

(optional background) sequences of functions

Depending on your background, you might have heard of:

pointwise convergence $f_n(x) \to f(x)$ (converges for every x) uniform convergence (convergence happens all at the same rate

Uniform convergence is stronger and has benefits - you can pass limits, derivatives, integrals, etc. through uniform convergence with no problem.

In this course we have sometimes magically passed these things along with the $E()$ operator through power series, because power series converge uniformly inside their radius of convergence.

But don't worry if you've never heard of this or forgot it all.

sequences of random variables

Very common in probability and statistics. We have seen some already:

As a model for a variable in a dataset we have considered the "i.i.d." sequence $X_1, X_2, \ldots, X_n$.

I have introduced the notion of "Sample Average" $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

When we derived Poisson from Binomial, we (implicitly) considered a sequence $X_n \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right)$ and wondered about $n \to \infty$.

We're going to wonder again about $n \to \infty$

Again, with random variables we care most about probabilities and not their actual values.

Case 1: getting closer to a constant, probably

Consider $X_1, X_2, \ldots, X_n$ i.i.d. with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, and consider $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

From last week:

$$E\left(\overline{X}_n\right) = \mu \qquad \text{and} \qquad \mathrm{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$$

What happens when n gets bigger?

simulation example 1 - Bernoulli(0.5)

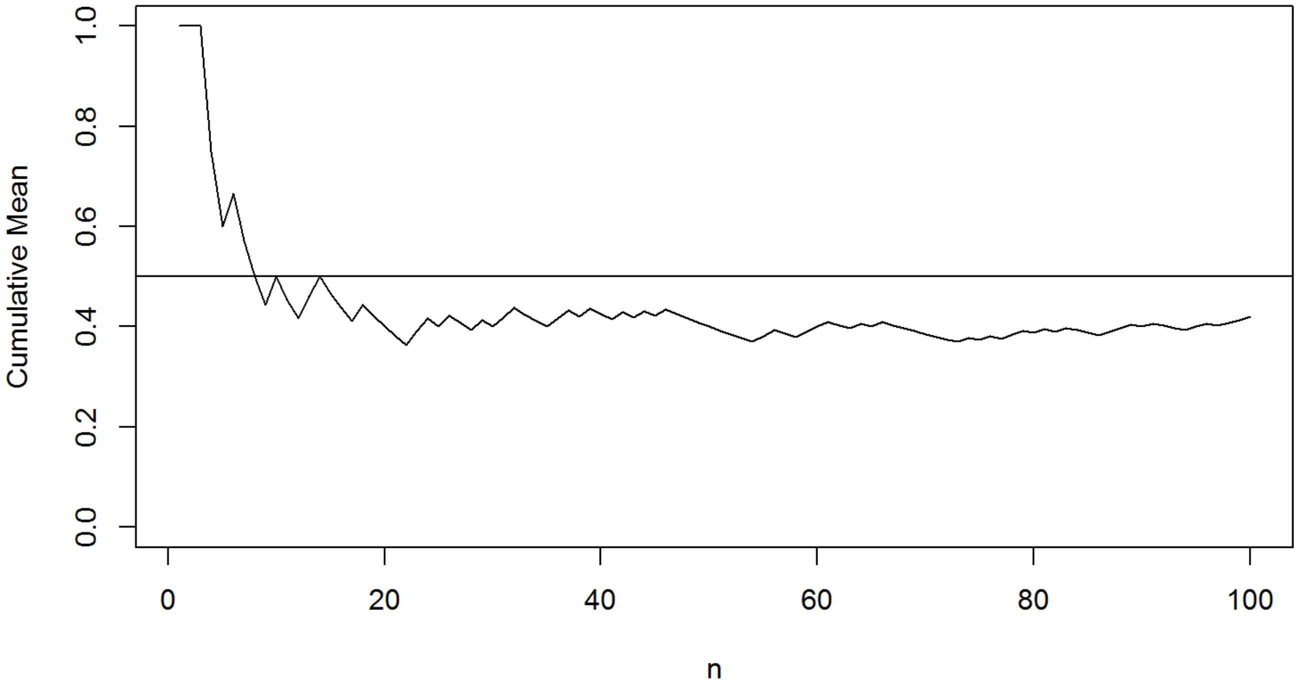
My computer can pretend to observe Bernoulli random variables. Here are $n = 30$ Binomial(1,0.5) simulations:

```
rbinom(n=30, size=1, prob=0.5)
```

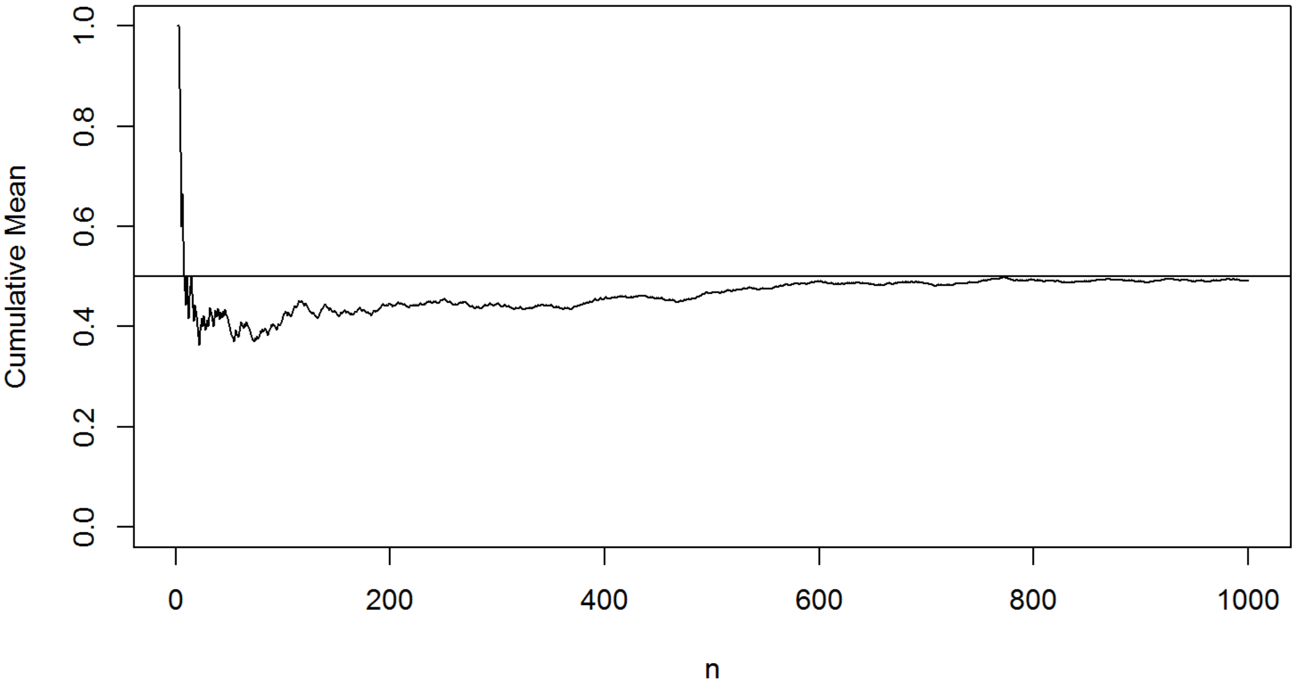
```
##  [1] 0 0 0 0 1 1 0 1 1 0 1 0 0 0 1 0 0 0 1 0 1 0 0 1 1 1 0 0 0 0
```

I am going to let n get larger and plot n versus cumulative values of $\overline{X}_n$.

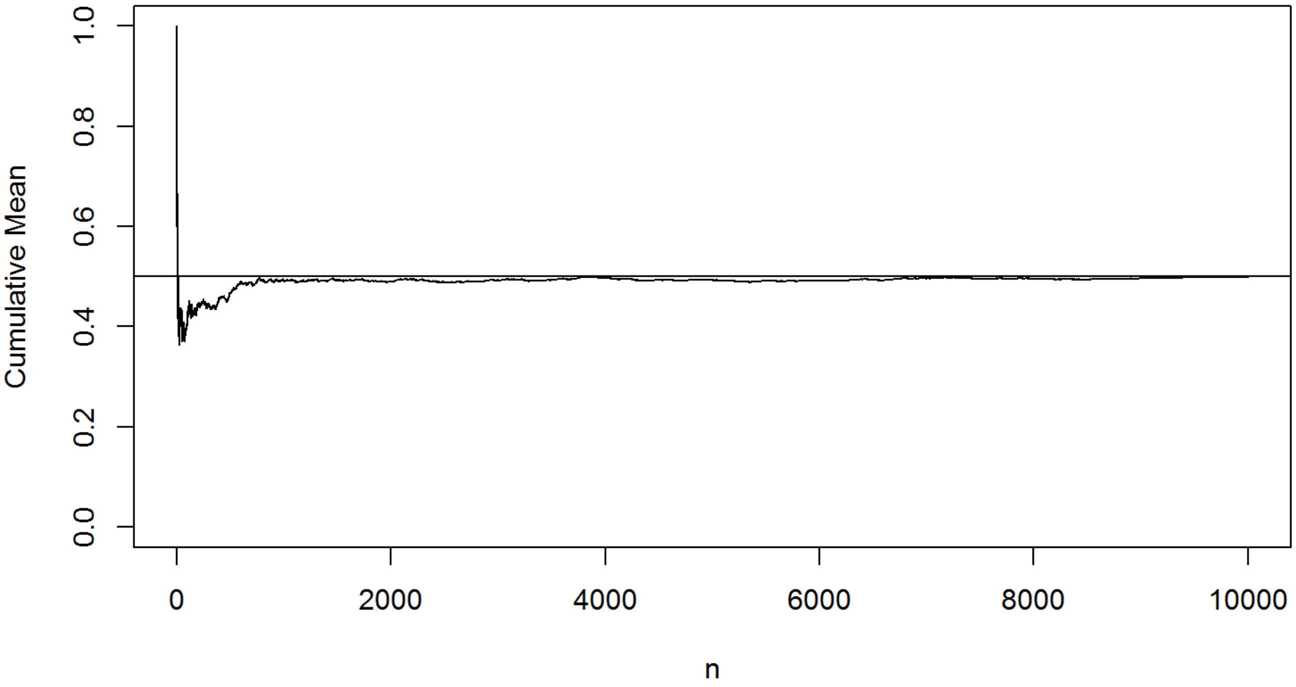
simulation example 1 - Bernoulli(0.5)



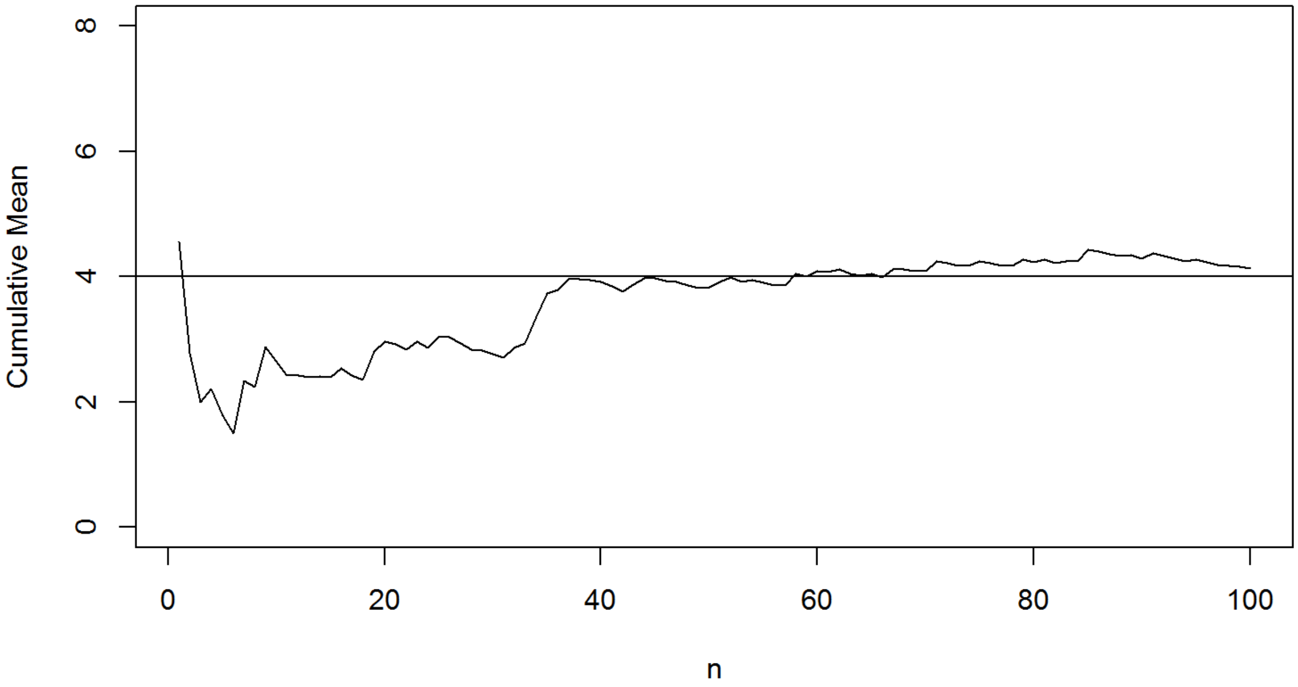
simulation example 1 - Bernoulli(0.5)



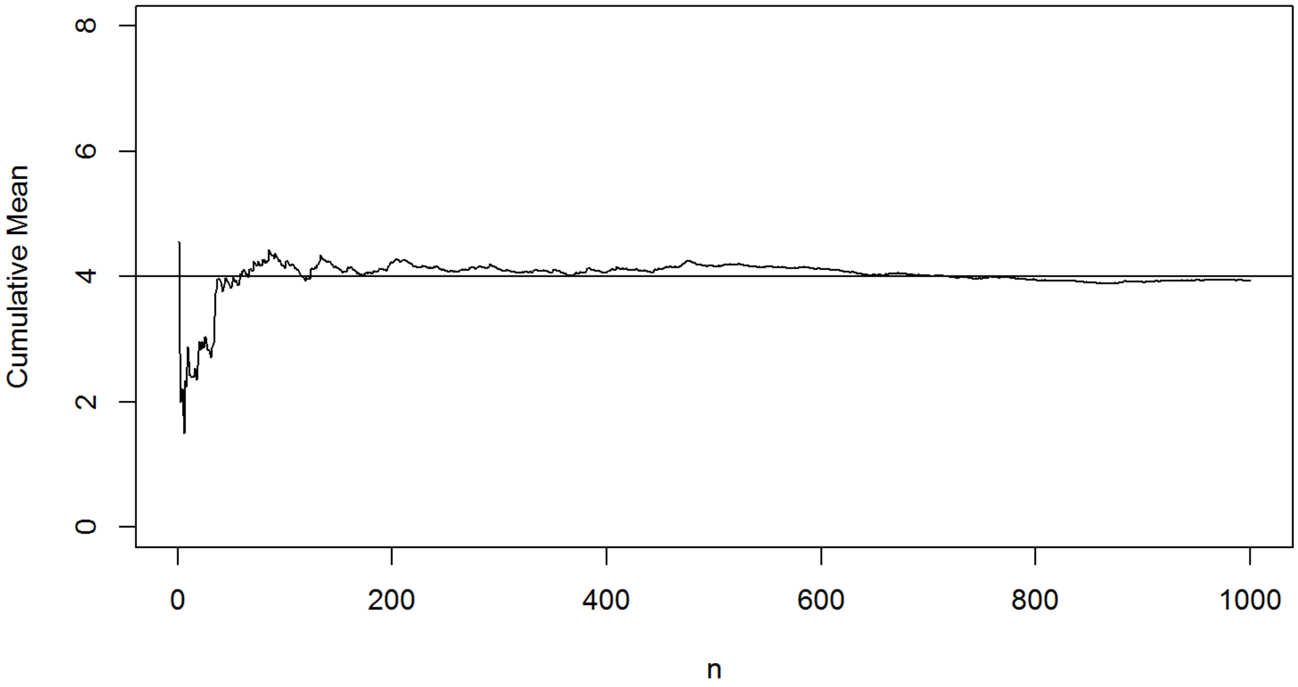
simulation example 1 - Bernoulli(0.5)



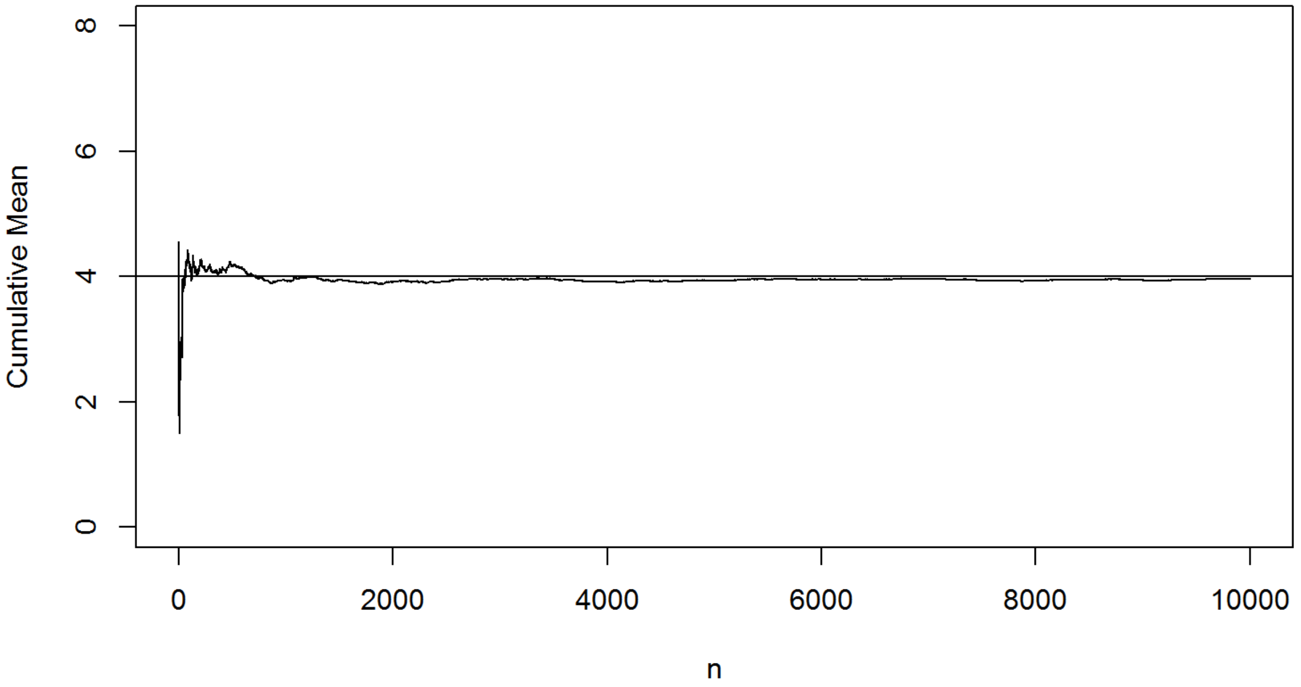
simulation example 2 - Exponential(0.25)



simulation example 2 - Exponential(0.25)



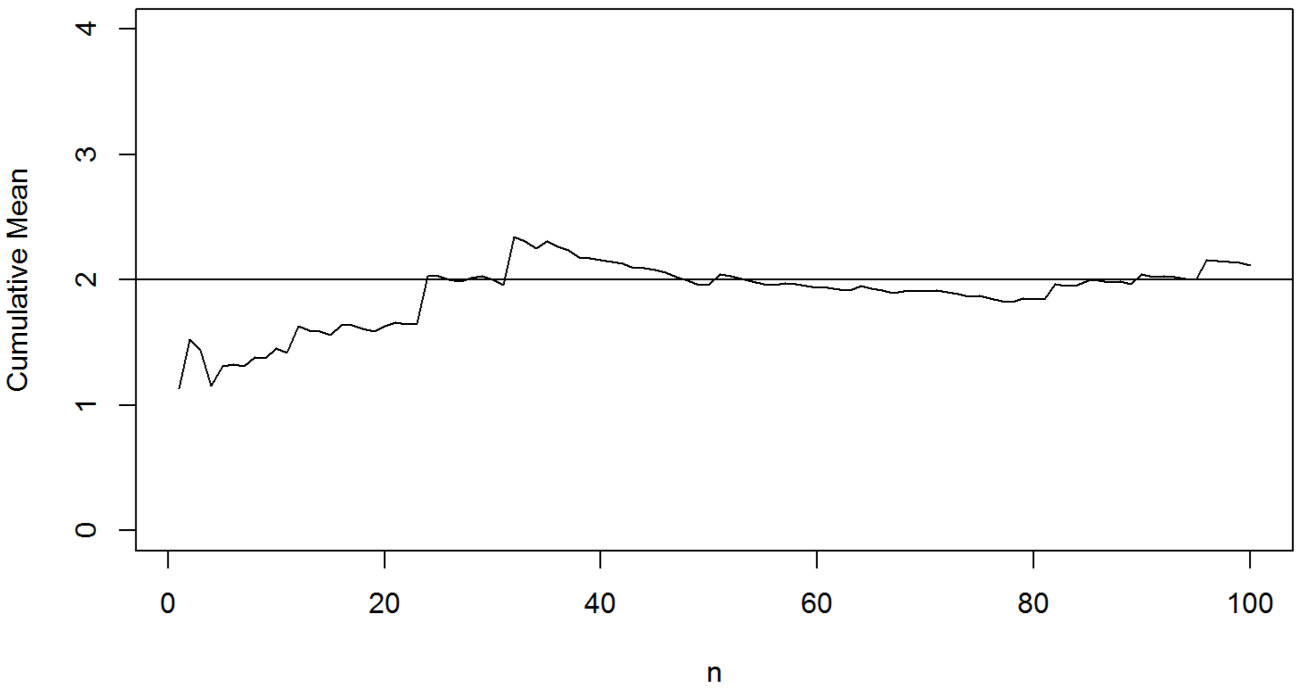
simulation example 2 - Exponential(0.25)



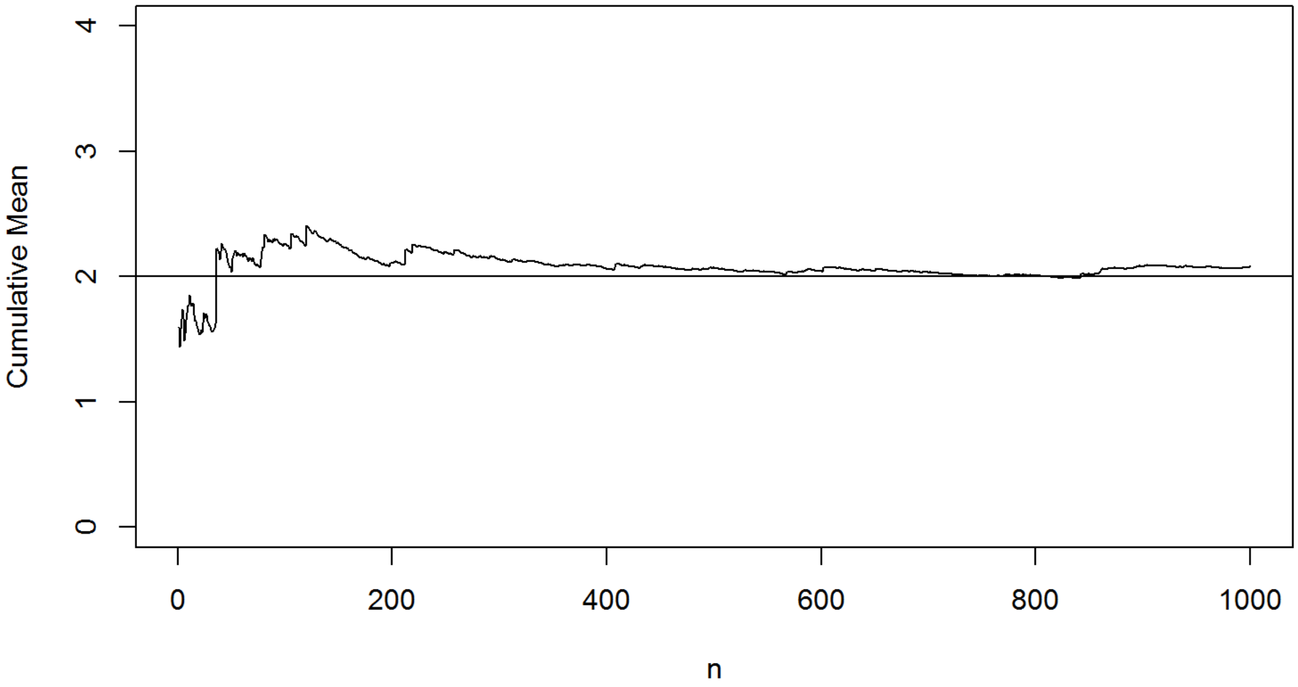
simulation example 3 - The Neil distribution

Positive random variable X with density $\frac{8}{\pi}\frac{x}{x^4+4}$

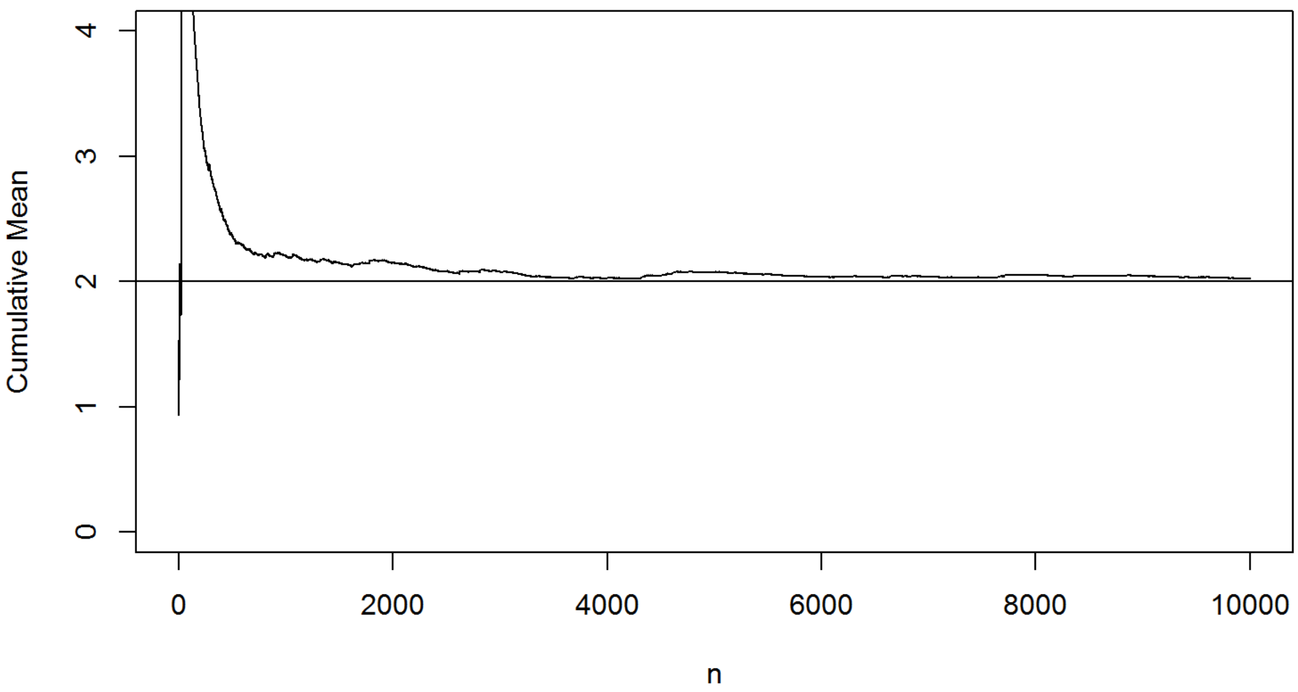
$E(X) = 2$ (hard!!) and $\mathrm{Var}(X)$ does not exist.



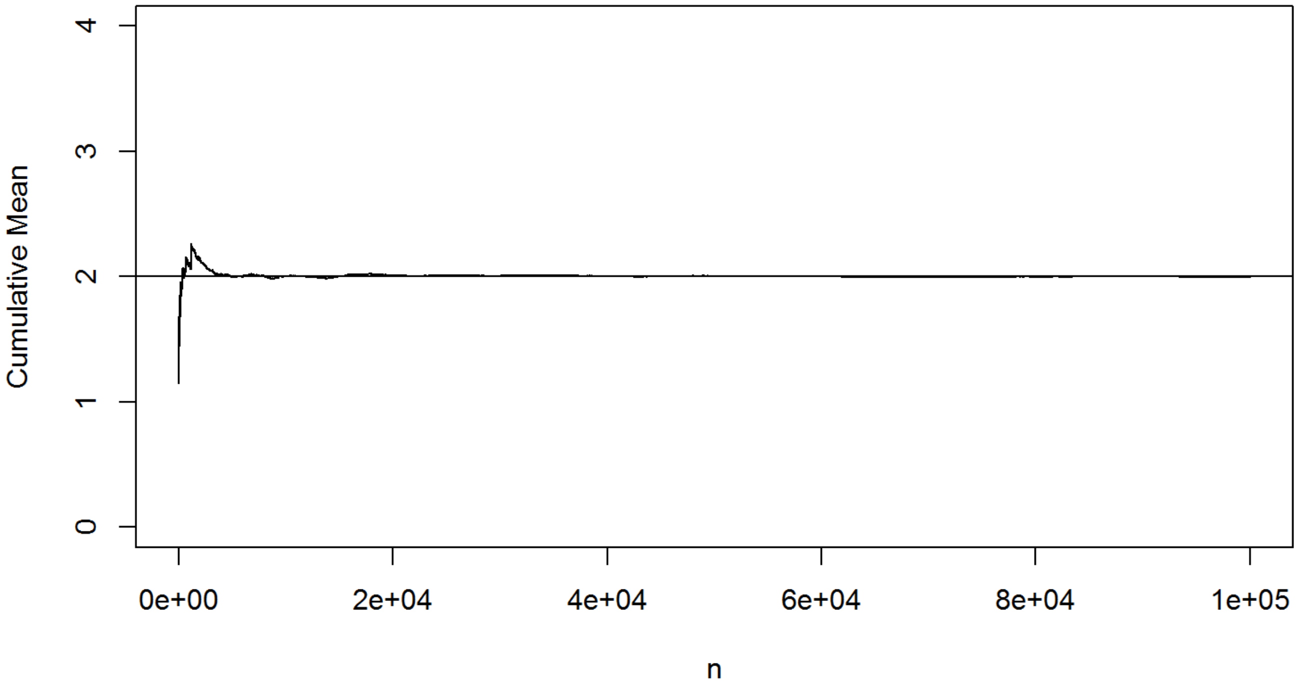
simulation example 3 - The Neil distribution



simulation example 3 - The Neil distribution

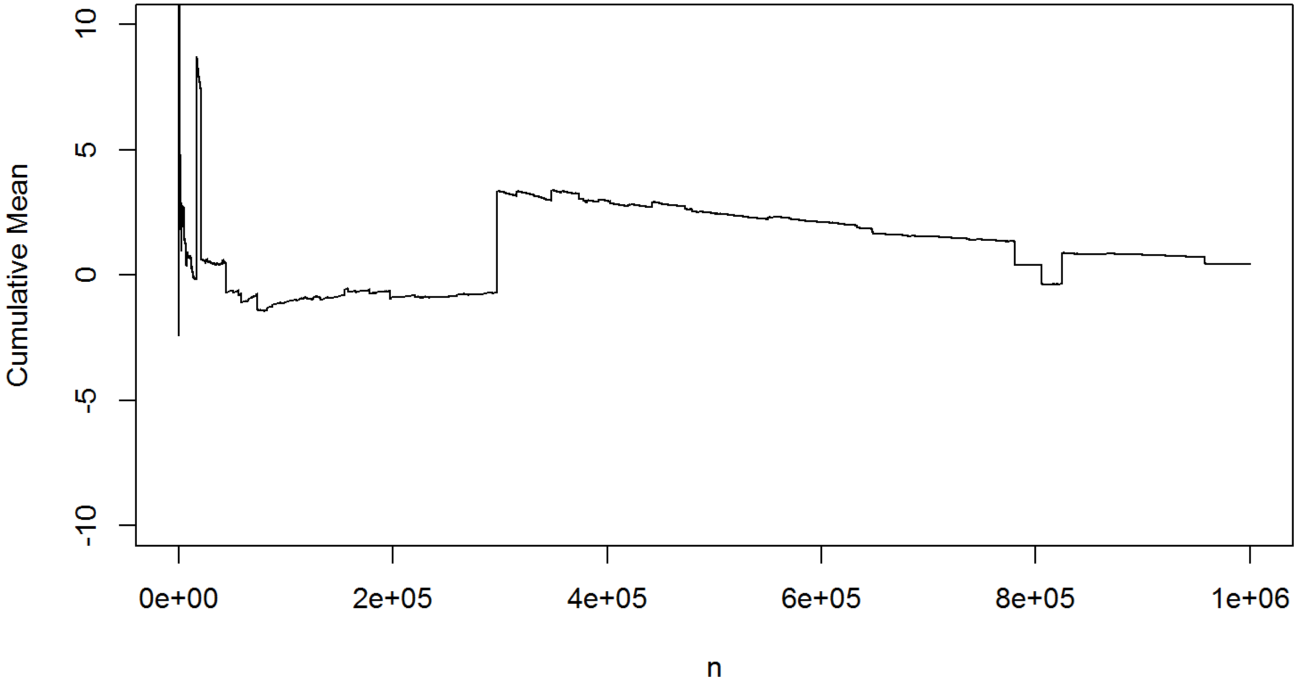


simulation example 3 - the Neil distribution



simulation example 4 - the Cauchy distribution

Density $\frac{1}{\pi}\frac{1}{1+x^2}$. No expected value



the (weak) Law of Large Numbers

Examples 1 and 2 had random variables with mean μ and variance σ^2. The convergence of the $\overline{X}_n$ to μ is explained by…

Theorem: If $X_1, X_2, X_3, \ldots$ are independent with the same mean μ and variance σ^2, then for all $\epsilon > 0$:

$$\lim_{n \to \infty} P\left(\left| \overline{X}_n - \mu \right| > \varepsilon \right) = 0.$$

Proof:…

This does not explain example 3 (no variance), and doesn't really explain 4 either.

Good for theory and philosophy. Not so good in practice (rate of convergence?)

convergence in probability

The WLLN is an example of convergence in probability:

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0,$$

and in the WLLN $X = \mu$ is a constant random variable.

Common notation: $X_n \xrightarrow{P} X$

Convergence in probability has to do with values of random variables.

The distribution of X (what we care about) only deals with things like $P(a < X < b)$, which is less stringent than convergence in probability.

convergence in distribution

Roughly speaking, the distribution of X_n converges to the distribution of X if their cdfs converge: $F_{X_n} \to F_X$

Formal definition: X_n converges in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at every point where $F_X(x)$ is continuous.

Common notations:

$$X_n \Longrightarrow X$$
$$X_n \xrightarrow{D} X$$

Definition not so easy to use.

verifying convergence in distribution

Stated without proof; all imply $X_n \xrightarrow{D} X$.

Theorem: X_n with pmf $p_{X_n}(x)$ that converge to a pmf $p_X(x)\dots$

Example: $X_n \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right)$ and $X \sim \text{Poisson}(\lambda)$

Theorem: X_n with density $f_{X_n}(x)$ that converge to a density $f_X(x)\dots$

Future example for STA261 students: $X_n \sim t_n$ and $X \sim N(0,1)$

THEOREM: X_n with m.g.f. $m_{X_n}(t)$ that converge to an m.g.f. $m_X(t)$ in a neighborhood of 0 implies $X_n \xrightarrow{D} X$.

Example: $X_n \sim \text{Binomial}\left(n, \frac{\lambda}{n}\right)\dots$

the fundamental theorem of statistics

the distribution of $\overline{X}_n$

Let's consider again $X_1, X_2, \ldots$ now i.i.d. with m.g.f. $m(t)$. The common mean and variance are μ and σ^2.

$E\left(\overline{X}_n\right) = \mu$ and $\text{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$

Consider:

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

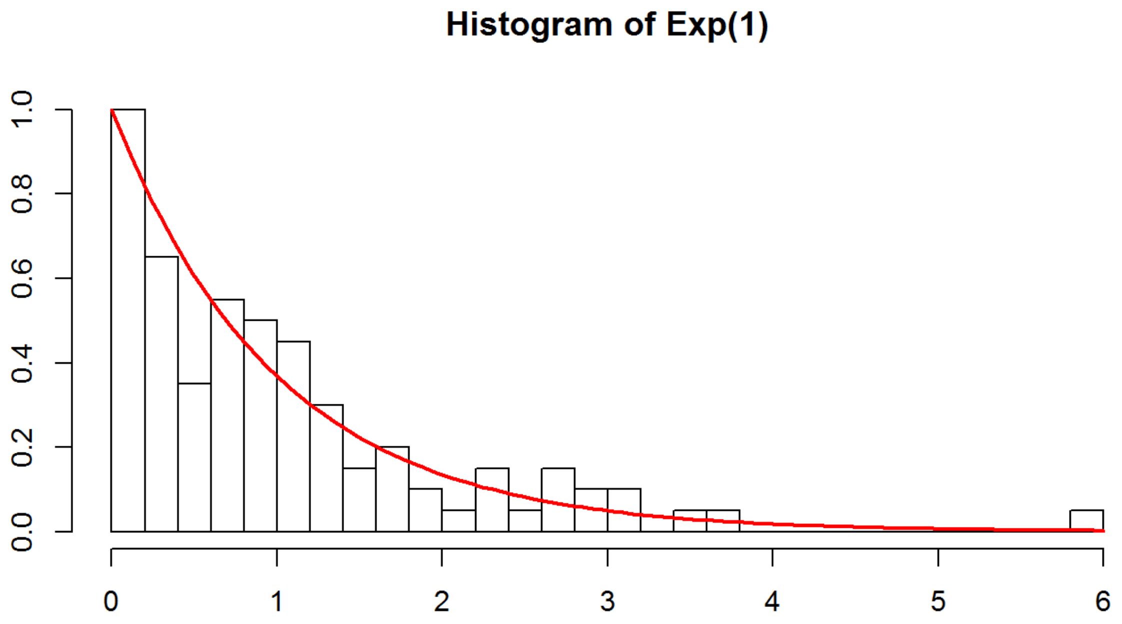
Then $E(Y_n) = 0$ and $\text{Var}(Y_n) = 1$.

What could be said about the distribution of Y_n?

histograms (for simulating Y_n)

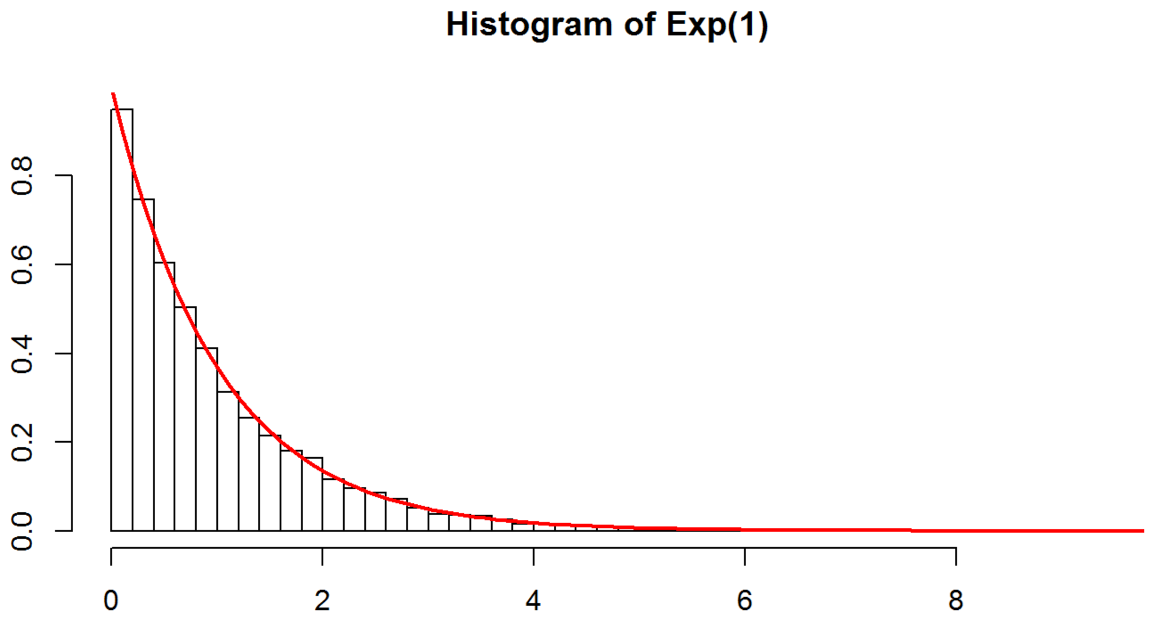
A histogram takes a sequence of numbers (the "data"), splits the range into "bins", and produces a bar graph of the count inside each bin.

A histogram is a "density estimator". Here's a histogram of $k = 100$ randomly samples from an $\mathrm{Exp}(1)$ distribution, with the density in red:



more simulation = smoother histogram

$k = 10^4$



simulating Y_n

Fix n and a distribtion for X_i.

1. Simulate a sample of size n from the distribution.

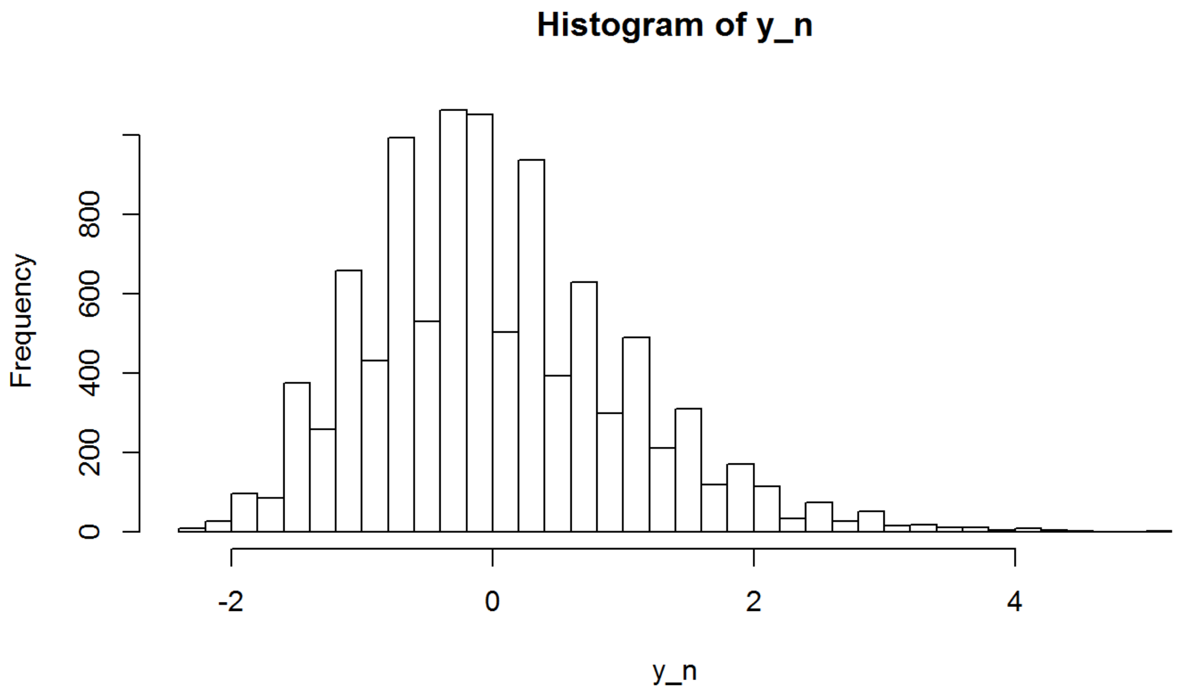
2. Calculate $Y_n^{(1)}$ from this sample.

3. Repeat k times to obtain $Y_n^{(1)}, \dots, Y_n^{(k)}$

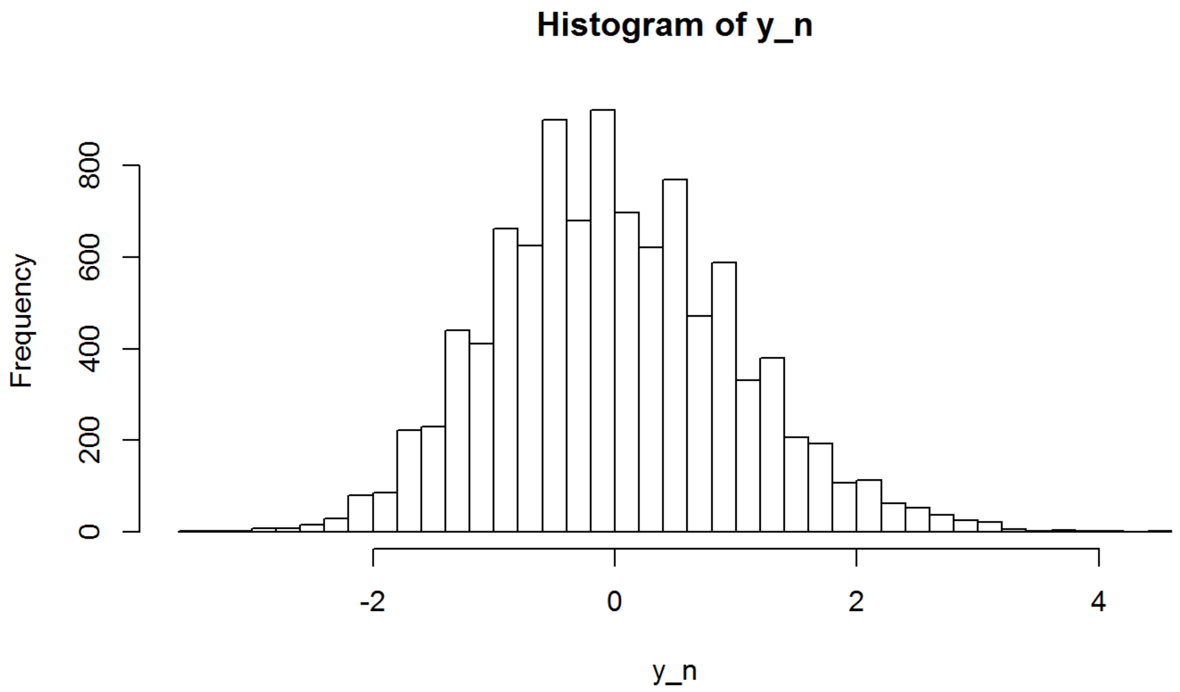
4. Make a histogram of the $Y_n^{(j)}$

Important: n is fixed and fundamental to the simulation. A larger k makes a nicer histogram and is more of a choice to make.

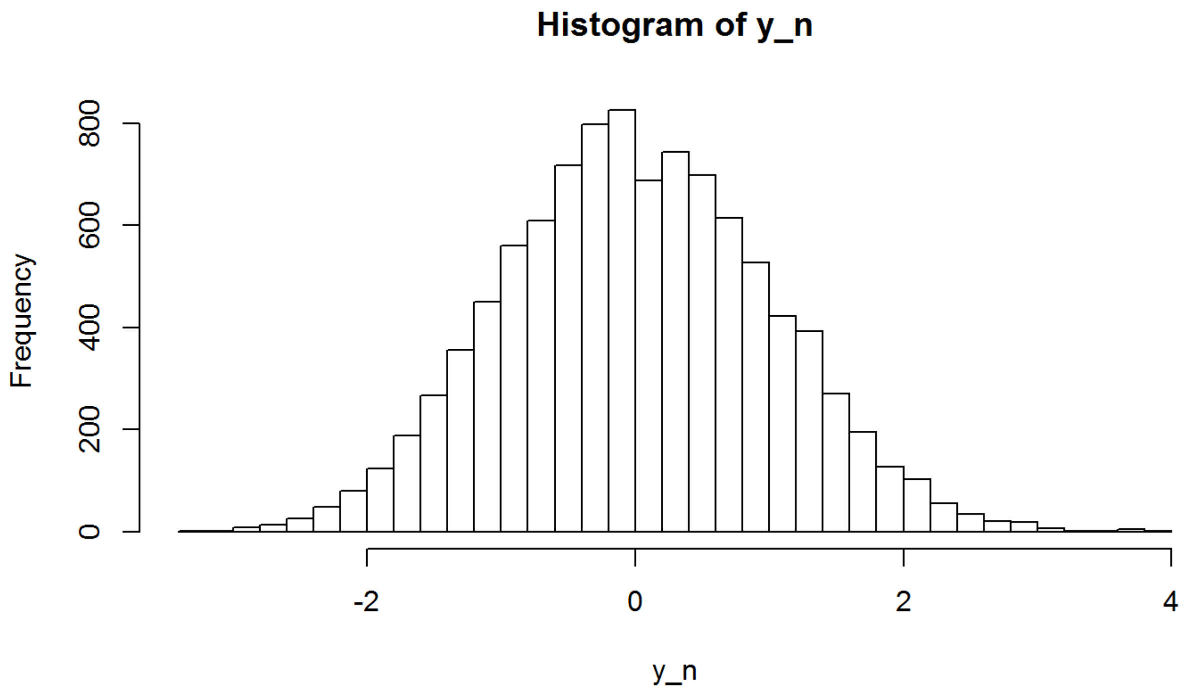
example 1 with $n = 10$ and $X_i \sim \text{Geometric}(1/3)$



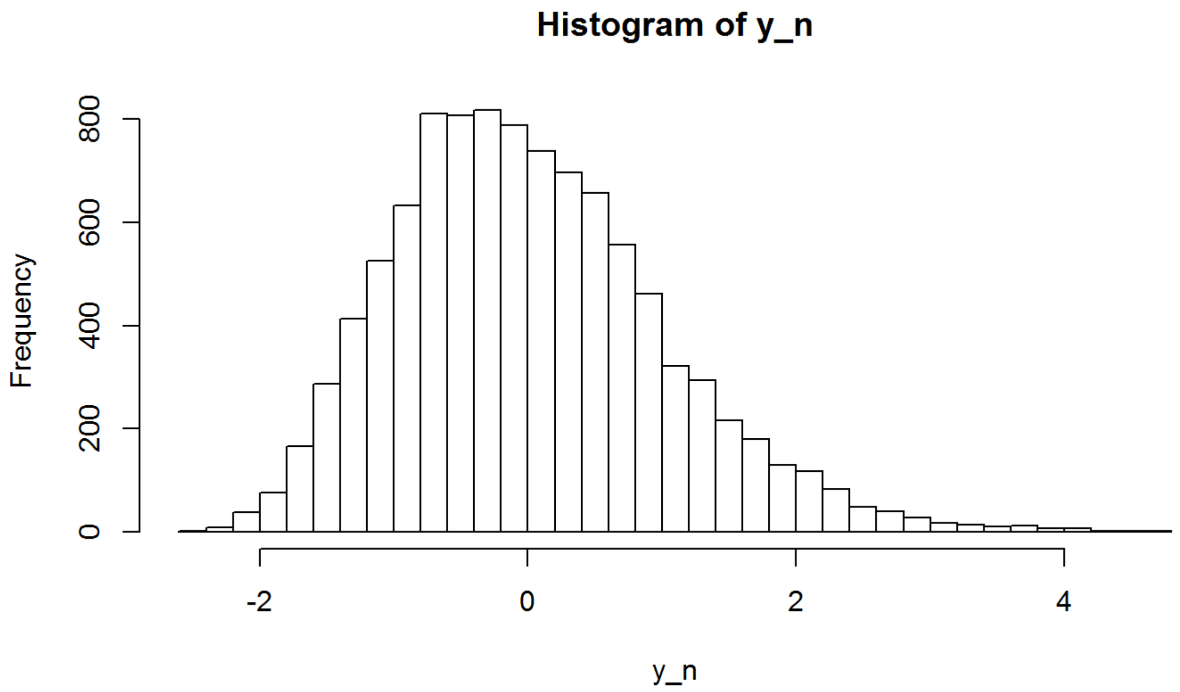
example 1 with $n = 50$ and $X_i \sim \text{Geometric}(1/3)$



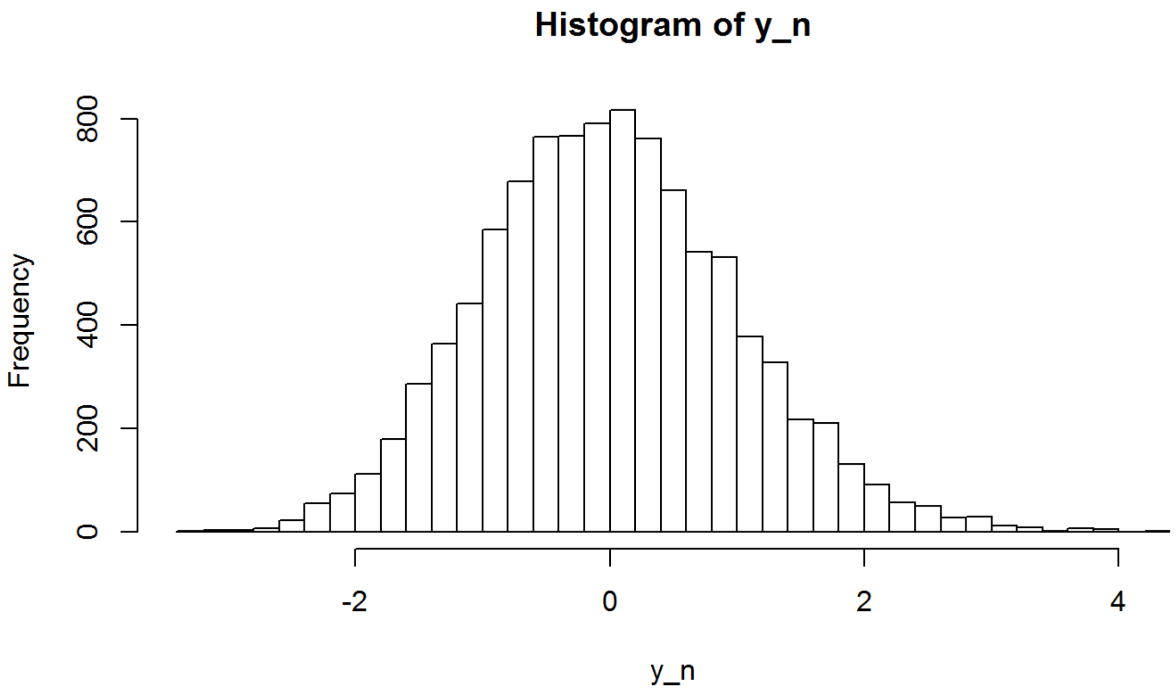
example 1 with $n = 500$ and $X_i \sim \text{Geometric}(1/3)$



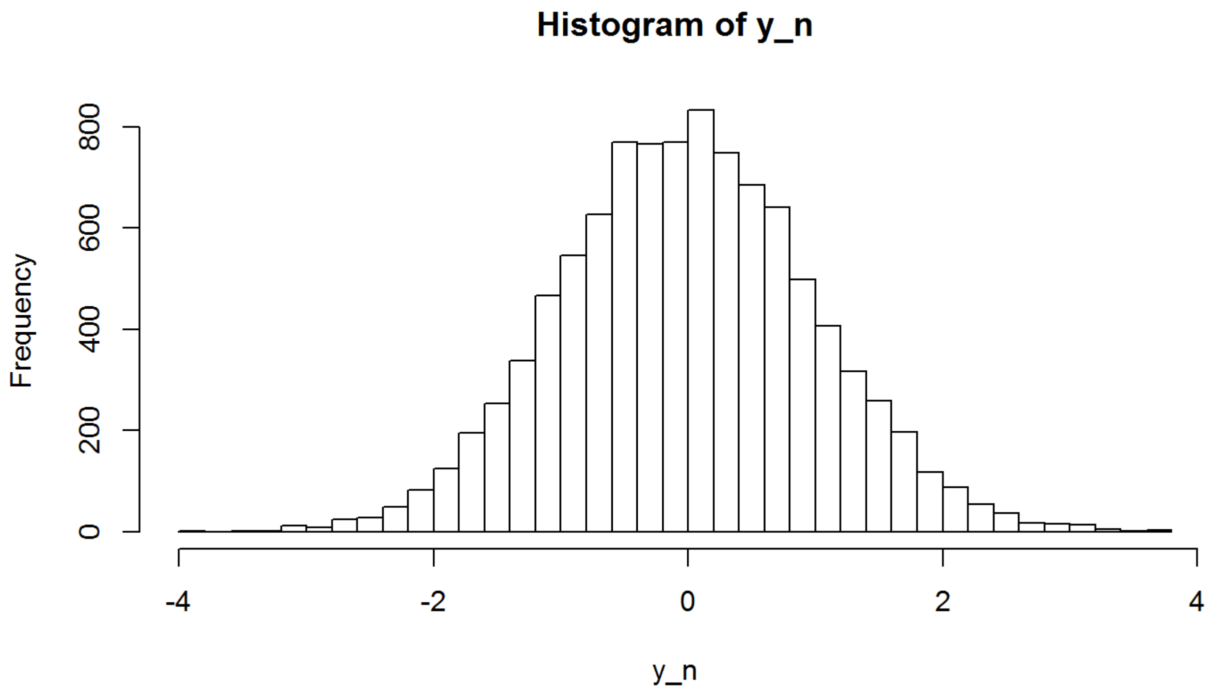
example 2 with $n = 10$ and $X_i \sim \mathrm{Exp}(0.25)$



example 2 with $n = 50$ and $X_i \sim \mathrm{Exp}(0.25)$



example 2 with $n = 500$ and $X_i \sim \text{Exp}(0.25)$



The Central Limit Theorem

Theorem: $X_1, X_2, \dots$ are i.i.d. with m.g.f. $m(t)$, mean μ, and variance σ^2. Then:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z$$

where $Z \sim N(0,1)$.

Proof:…

This is a limit theorem. What is crucial is that the convergence can be fast.

normal approximations

For $X_1, X_2, \ldots, X_n$ with mean μ and variance σ^2 and n large enough

$$\sum_{i=1}^{n} X_i \sim^{\text{approx}} N(n\mu, n\sigma^2)$$

$$\overline{X} \sim^{\text{approx}} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim^{\text{approx}} N(0, 1)$$

depends on the underlying distribution. The more "skewed" or "heavy-tailed", the larger the n required.

example - Uniform(0,1)

$X_1, X_2, \ldots, X_n$ are Uniform(0,1) and $n = 20$. What's the chance that $\sum X_i > 11$