STA286 All Notes 2017

Neil Montgomery

admin

contact, notes

date format	YYYY-MM-DD – *All Hail ISO8601!!!*
instructor	Neil Montgomery
email	`neilm@mie.utoronto.ca`
office	BA8137
office hours	W11-1
website	portal (announcements, grades, suggested exercises, etc.)
github	`https://github.com/sta286-winter-2017` (lecture material, code, etc.)

Lecture notes and other course timing matters will be organized by *lecture number* and not *lecture date*, due to two lecture sections.

evaluation, book, tutorials

what	when	how much
midterm 1	2017-02-13	20%
midterm 2	2017-03-27	30%
exam	TBA	50%

The book is Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K., 2012. *Probability & statistics for engineers & scientists.* 9th edition.

I will suggest exercises from this book each week. Your TA will work through some of them in tutorial each week.

Tutorials start TBA.

software

The course begins and ends with data analysis, with a long stretch of probability theory in the middle.

Data analysis requires a computer. Also, some concepts can be illustrated using simulation, which also requires a computer.

We will be using `R`. It's pretty good at data analysis.

language	interpreter	integrated development environment
`R`	`R`	`RStudio`

Some detailed instructions and suggestions for installation and configuration appear on the course website.

I will try to impart some data analysis workflow wisdom throughout the course. Some already appears in the detailed instructions.

MATLAB SUCKS!

what is a dataset?

most datasets are rectangles

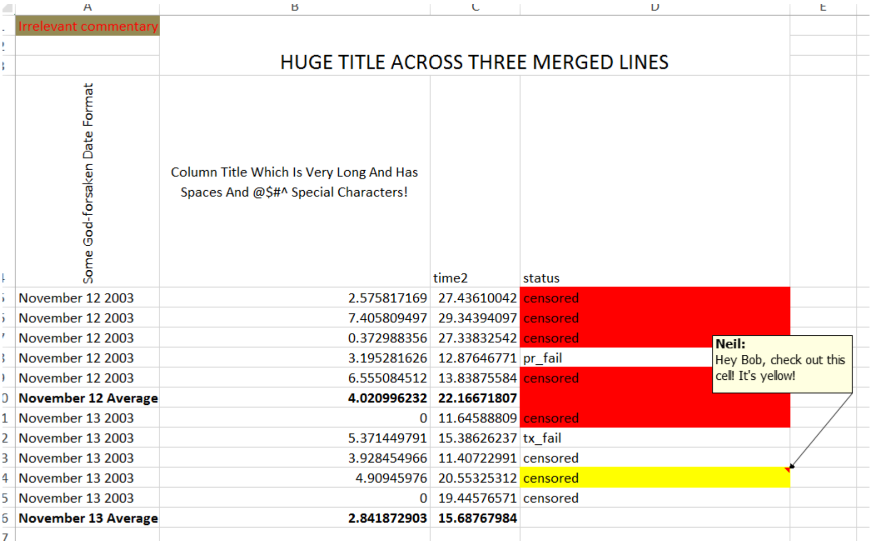
Columns are the *variables*.

The top row has the names of the variables; possibly chosen wisely.

Rows are the *observations* of measurements taken on *units*.

There are no averages, no comments (unless in a “comment” variable), no colors, no formatting, no plots, no capes!

not a dataset

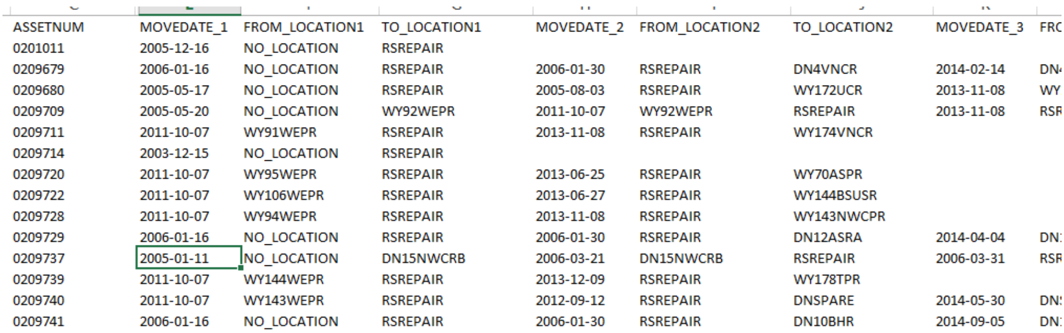


	A	B	C	D	E
1	Irrelevant commentary				
2					
3		HUGE TITLE ACROSS THREE MERGED LINES			
4	Some God-forsaken Date Format	Column Title Which Is Very Long And Has Spaces And @$#^ Special Characters!	time2	status	
5	November 12 2003	2.575817169	27.43610042	censored	
6	November 12 2003	7.405809497	29.34394097	censored	
7	November 12 2003	0.372988356	27.33832542	censored	
8	November 12 2003	3.195281626	12.87646771	pr_fail	
9	November 12 2003	6.555084512	13.83875584	censored	
10	**November 12 Average**	**4.020996232**	**22.16671807**		
11	November 13 2003	0	11.64588809	censored	
12	November 13 2003	5.371449791	15.38626237	tx_fail	
13	November 13 2003	3.928454966	11.40722991	censored	
14	November 13 2003	4.90945976	20.55325312	censored	
15	November 13 2003	0	19.44576571	censored	
16	**November 13 Average**	**2.841872903**	**15.68767984**		
17					

Neil:
Hey Bob, check out this cell! It's yellow!

Figure 1:

not a dataset



ASSETNUM	MOVEDATE_1	FROM_LOCATION1	TO_LOCATION1	MOVEDATE_2	FROM_LOCATION2	TO_LOCATION2	MOVEDATE_3	FRC
0201011	2005-12-16	NO_LOCATION	RSREPAIR					
0209679	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN4VNCR	2014-02-14	DN
0209680	2005-05-17	NO_LOCATION	RSREPAIR	2005-08-03	RSREPAIR	WY172UCR	2013-11-08	WY
0209709	2005-05-20	NO_LOCATION	WY92WEPR	2011-10-07	WY92WEPR	RSREPAIR	2013-11-08	RSR
0209711	2011-10-07	WY91WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY174VNCR		
0209714	2003-12-15	NO_LOCATION	RSREPAIR					
0209720	2011-10-07	WY95WEPR	RSREPAIR	2013-06-25	RSREPAIR	WY70ASPR		
0209722	2011-10-07	WY106WEPR	RSREPAIR	2013-06-27	RSREPAIR	WY144BSUSR		
0209728	2011-10-07	WY94WEPR	RSREPAIR	2013-11-08	RSREPAIR	WY143NWCPR		
0209729	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN12ASRA	2014-04-04	DN
0209737	2005-01-11	NO_LOCATION	DN15NWCRB	2006-03-21	DN15NWCRB	RSREPAIR	2006-03-31	RSR
0209739	2011-10-07	WY144WEPR	RSREPAIR	2013-12-09	RSREPAIR	WY178TPR		
0209740	2011-10-07	WY143WEPR	RSREPAIR	2012-09-12	RSREPAIR	DNSPARE	2014-05-30	DN
0209741	2006-01-16	NO_LOCATION	RSREPAIR	2006-01-30	RSREPAIR	DN10BHR	2014-09-05	DN

Figure 2:

an oil readings dataset (wide version)

```
## # A tibble: 612 × 17
##     Ident                Date WorkingAge   TakenBy    Fe     Al    Cu
##     <chr>              <dttm>      <dbl>     <chr> <dbl> <dbl> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings with Ident and TakenBy properly treated

```
## # A tibble: 612 × 17
##      Ident                Date WorkingAge   TakenBy    Fe    Al    Cu
##     <fctr>              <dttm>      <dbl>    <fctr> <dbl> <dbl> <dbl>
## 1   448576 1999-05-10 19:00:00        243 EMPL_0917    13     5    14
## 2   448576 1999-07-26 19:00:00        569 EMPL_0917    18     6    25
## 3   448576 1999-09-29 19:00:00        830 EMPL_9375    26     6    35
## 4   448576 1999-10-08 19:00:00        862 EMPL_0917    15     9    14
## 5   448576 1999-11-02 19:00:00        946 EMPL_9375    14     4    19
## 6   448576 1999-12-09 19:00:00       1088 EMPL_0917    18     5    23
## 7   448576 1999-12-27 19:00:00       1157 EMPL_9375    24     8    25
## 8   448576 2000-01-14 19:00:00       1238 EMPL_9375    27     9    34
## 9   448576 2000-02-15 19:00:00       1376 EMPL_9375    16     8    17
## 10  448576 2000-03-11 19:00:00       1492 EMPL_0917    20     8    20
## # ... with 602 more rows, and 10 more variables: Cr <dbl>, Si <dbl>,
## #   Pb <dbl>, Ph <dbl>, Ca <dbl>, Zn <dbl>, Mg <dbl>, Mo <dbl>,
## #   Sn <dbl>, Na <dbl>
```

oil readings dataset (long version)

```
## # A tibble: 7,956 × 6
##     Ident                Date WorkingAge   TakenBy element   ppm
##    <fctr>              <dttm>      <dbl>    <fctr>   <chr> <dbl>
## 1  448576 1999-05-10 19:00:00        243 EMPL_0917      Fe    13
## 2  448576 1999-07-26 19:00:00        569 EMPL_0917      Fe    18
## 3  448576 1999-09-29 19:00:00        830 EMPL_9375      Fe    26
## 4  448576 1999-10-08 19:00:00        862 EMPL_0917      Fe    15
## 5  448576 1999-11-02 19:00:00        946 EMPL_9375      Fe    14
## 6  448576 1999-12-09 19:00:00       1088 EMPL_0917      Fe    18
## 7  448576 1999-12-27 19:00:00       1157 EMPL_9375      Fe    24
## 8  448576 2000-01-14 19:00:00       1238 EMPL_9375      Fe    27
## 9  448576 2000-02-15 19:00:00       1376 EMPL_9375      Fe    16
## 10 448576 2000-03-11 19:00:00       1492 EMPL_0917      Fe    20
## # ... with 7,946 more rows
```

the main questions

- where did the data come from?
  - were the units chosed randomly from a population?
  - were the units randomly assigned into groups?
- what are the (joint) *distributions* of the data?

random sample, experiment, observational data

Sometimes the data come from a *random sample* from a larger *population*, in which case statements about the sample can apply to the population using laws of probability.

(Not a focus of this course.)

Sometimes data come from an *experiment* where units are randomly assigned to different *levels* of one or more *factors*, in which cause cause-and-effect can be inferred using laws of probability.

Often the data are just some records of what happened. Grander inferences might be made, but only on a subject-matter basis.

distribution (informally)

- A *distribution* is a
  - Complete description of...
  - ...the possible values of one or more variables...
  - ...and the relative frequency of those values.
- A dataset contains **empirical** information about distribution(s) that can be assessed
  - numerically
  - graphically

through a process called *exploratory data analysis*

a taxonomy of variables

- Numerical or categorical?
    - Numerical: length, ppm, time-to-event, etc.
    - Categorical: yes/no, colour, etc.
    - Lots of grey areas even in this classification!
        - Categories can have an inherent order
        - "Likert scale" (strongly disagree coded as 1 and so on. . .)
- Numerical variables could be discrete (counting something) or continuously measured.

numerical summaries of dataset variables — definitions first with examples after

sample measures of “location”

The dataset is often called the “sample” (no matter where the data came from).

For a particular numerical variable in the sample with observations:

$$\{x_1, x_2, \ldots, x_n\}$$

the *sample average* is just the arithmetic mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Could be sensitive to extreme observations.

sample medians, sample percentiles

Order the observations:

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$$

A number that divides the observations into two groups is called a *sample median*. For example:

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & : n \text{ odd} \\ \left(x_{(n/2)} + x_{(n/2+1)} \right) /2 & : n \text{ even} \end{cases},$$

which is harder to write out than it is to understand.

A *sample p^{th} percentile* has p% of the data below or equal to it. Special cases include (sample. . .): quartiles, quintiles, deciles, and indeed the median itself.

sample measures of variation of a numerical variable

Very (too?) simple measure: *sample range* which is just $x_{(n)} - x_{(1)}$.

More common to consider the set of deviations from the sample mean:

$$x_i - \overline{x}$$

Adding them up just gives 0, so instead consider positive functions such as:

$$|x_i - \overline{x}| \qquad \text{or} \qquad (x_i - \overline{x})^2$$

Summing up over all the observations gives the *sum of absolute deviations* (aka SAD) and the *sample variance* respectively. Notation and formula:

$$s^2 = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

sample standard deviation

s^2 is essentially the average squared deviation. (More on $n - 1$ later in the course.)

The sample variance is good for theory but has an inconvenient unit. More practical is the *sample standard deviation*:

$$s = \sqrt{s^2}$$

numerical summaries for categorical variables

The oil readings data had one categorical variable, the Ident variable which is just a serial number. I added a fake one TakenBy for illustration.

```
## # A tibble: 5 × 17
##    Ident       Date WorkingAge   TakenBy    Fe    Al    Cu    Cr
##   <fctr>     <date>      <dbl>    <fctr> <dbl> <dbl> <dbl> <dbl>
## 1 448576 1999-05-10        243 EMPL_0917    13     5    14     1
## 2 448576 1999-07-26        569 EMPL_0917    18     6    25     1
## 3 448576 1999-09-29        830 EMPL_9375    26     6    35     1
## 4 448576 1999-10-08        862 EMPL_0917    15     9    14     1
## 5 448576 1999-11-02        946 EMPL_9375    14     4    19     1
## # ... with 9 more variables: Si <dbl>, Pb <dbl>, Ph <dbl>, Ca <dbl>,
## #   Zn <dbl>, Mg <dbl>, Mo <dbl>, Sn <dbl>, Na <dbl>
```

tables of counts (or proportions)

A categorical variable could also be called a *factor* variable with *levels*, and to tabulate the frequency of each level is the way to summarize.

```
## # A tibble: 25 × 3
##      Ident     n proportion
##     <fctr> <int>      <dbl>
## 1   448572    31 0.05065359
## 2   448574    31 0.05065359
## 3   448576    36 0.05882353
## 4   448577    29 0.04738562
## 5   448578    34 0.05555556
## 6   448579    36 0.05882353
## 7   448580    28 0.04575163
## 8   448581    31 0.05065359
## 9   448582    41 0.06699346
## 10 448583    42 0.06862745
## # ... with 15 more rows
```

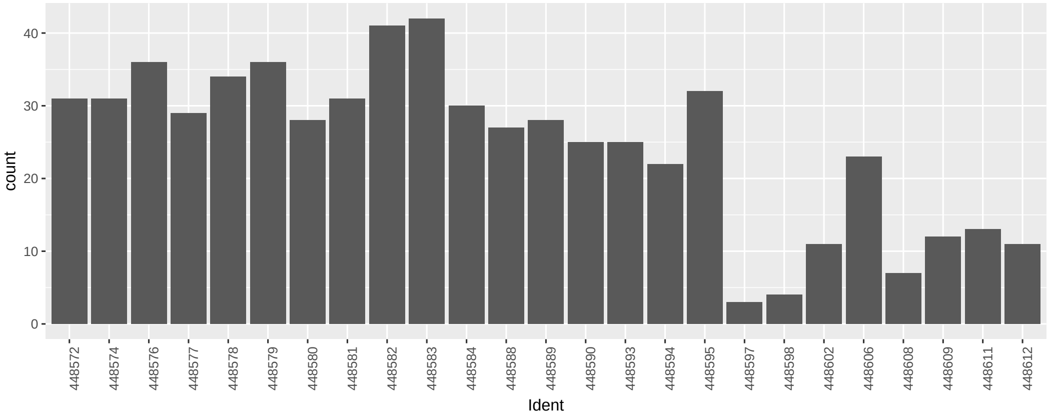
two-way classification with Ident and TakenBy

```
##            Ident
## TakenBy     448572 448574 448576 448577 448578 448579 448580 448581 4485
##   EMPL_0592     18     16      0      0     12      0      0      7
##   EMPL_0917      0      0     18     11      0     22     10      0
##   EMPL_2095      8      8      0      0      8      0      0      7
##   EMPL_4925      0      0     10      9      0      6     10      0
##   EMPL_9134      5      7      0      0     14      0      0     17
##   EMPL_9375      0      0      8      9      0      8      8      0
##            Ident
## TakenBy     448583 448584 448588 448589 448590 448593 448594 448595 4485
##   EMPL_0592      0      0      0     10      0     10      0      0
##   EMPL_0917     24      9     11      0     13      0     10     18
##   EMPL_2095      0      0      0     12      0      7      0      0
##   EMPL_4925     10     11     11      0      7      0     10     10
##   EMPL_9134      0      0      0      6      0      8      0      0
##   EMPL_9375      8     10      5      0      5      0      2      4
##            Ident
```

graphical summaries

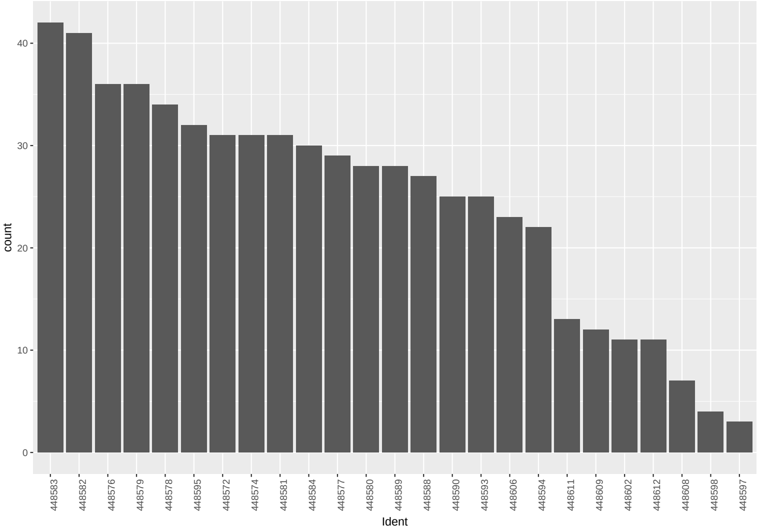
barchart

A barchart is a table of counts, in graphical form.

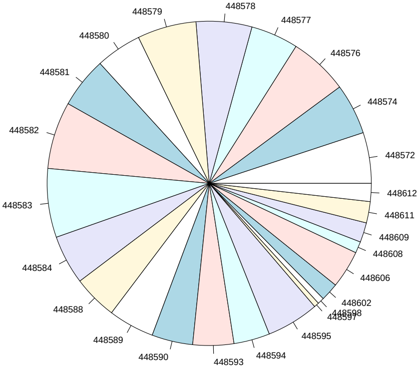


“Pareto” chart

Ordered by count.



piecharts are problematic

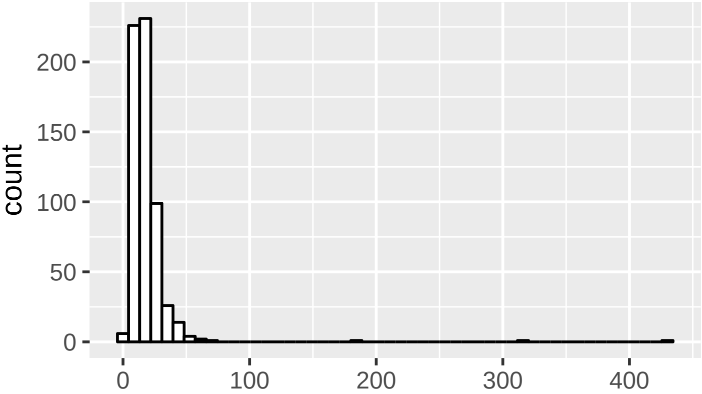


histograms

A histogram is a special case of a barchart.

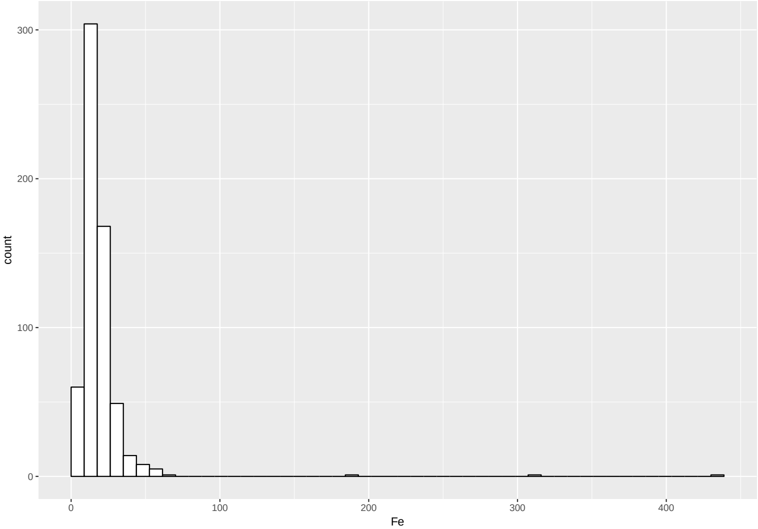
A numerical variable is split into classes and a barchart is made from the table of counts of obvservations within each class.

Histograms are done by the computer. Always play around with the number of classes.

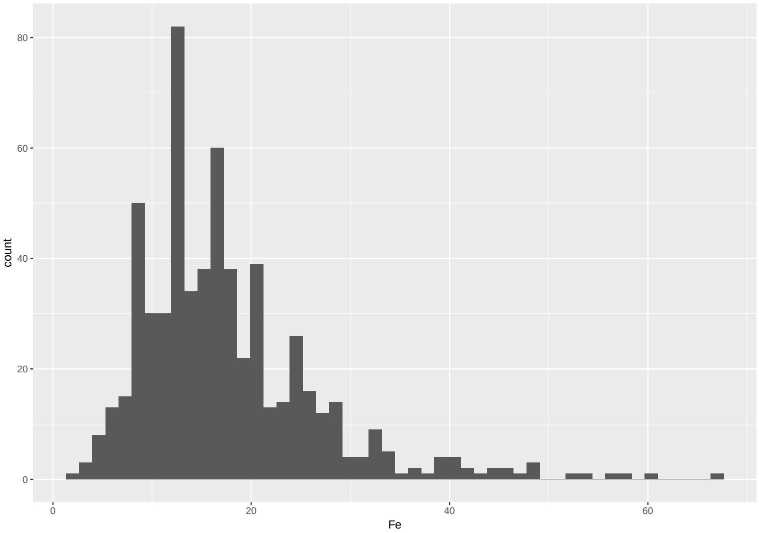


histograms are hard to implement!

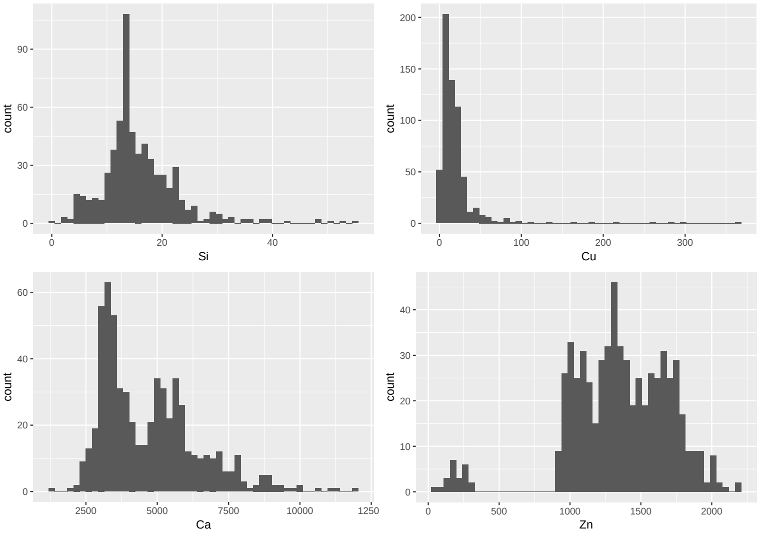
Better picture around 0. Possibly not important for EDA?



histogram without those really big values

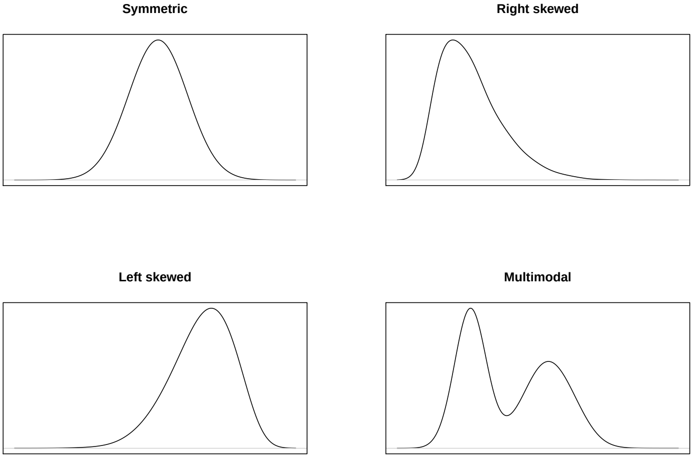


a few more ppm histograms



“shapes” of “distributions”

To use a histogram, *glance* at it and look for any of the following (without getting fooled by plot artefacts):



Symmetric

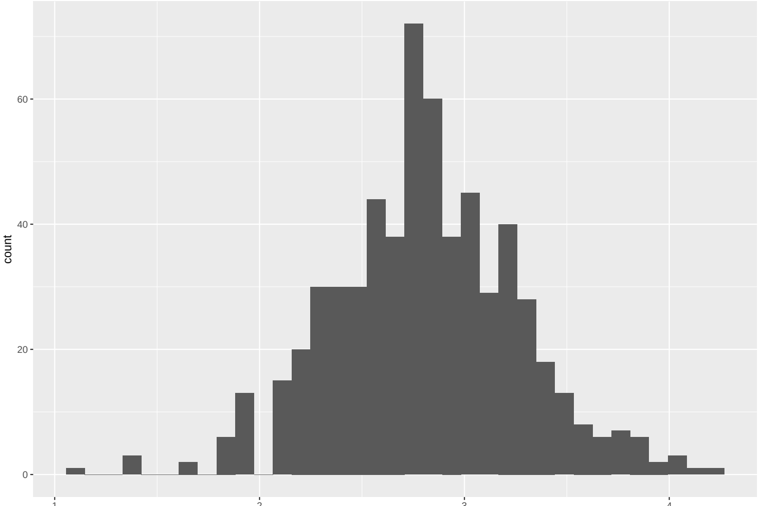
Right skewed

Left skewed

Multimodal

transforming variables

Apply log or square root to a variable will change the shape of the empirical distribution, e.g. transform right-skewed to symmetric.



boxplots

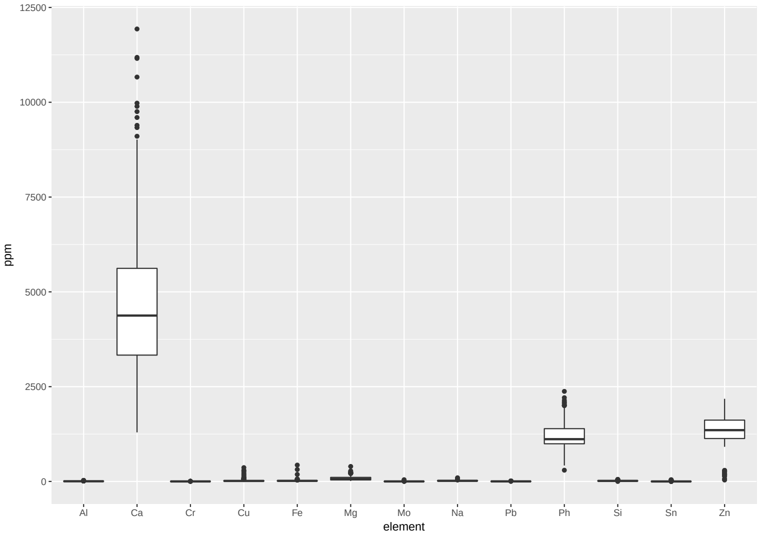
A special plot of these (or similar) five numbers:

min 25th percentile median 75th percentile max

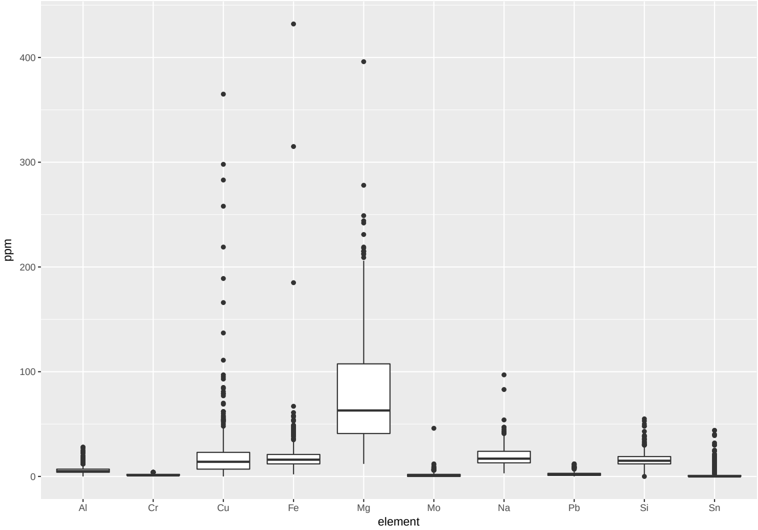
is called a *boxplot*. Often the extreme values are shown individually (see documentation for the (irrelevant) details.)

Best as *side-by-side* boxplots with more than one varaible on the same scale.

boxplot example - I

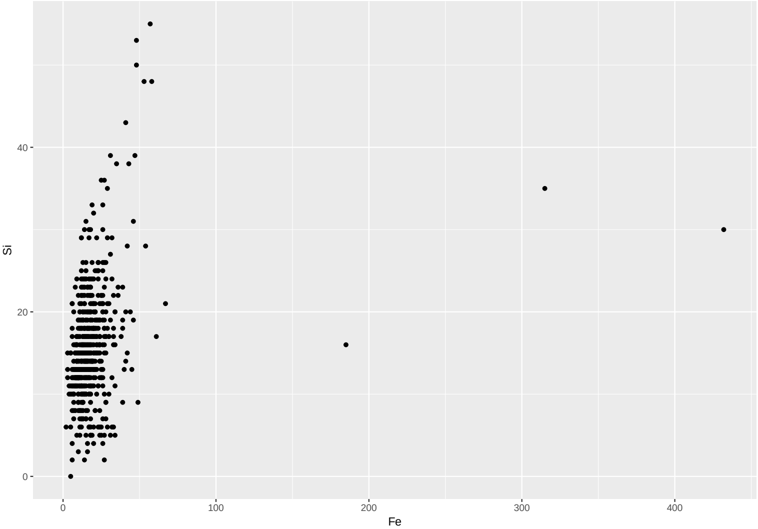


boxplot example - II

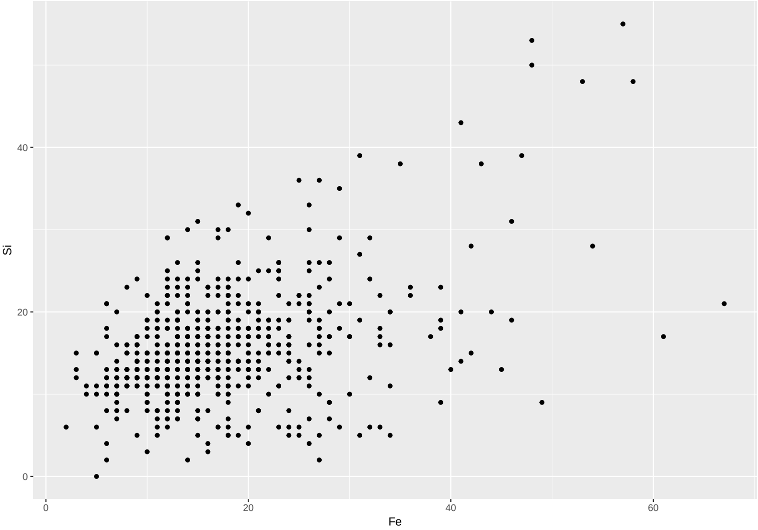


scatterplot

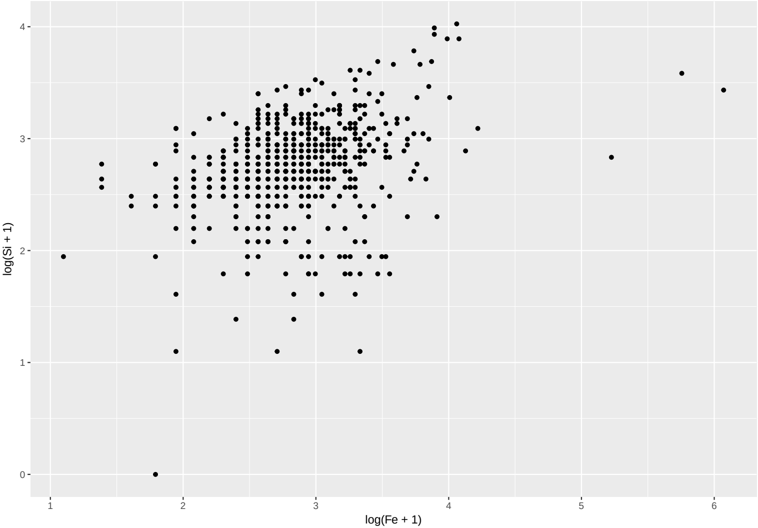
A graphic for two numerical variables, e.g. `Fe` and `Si`



Fe vs Si without the "outliers"

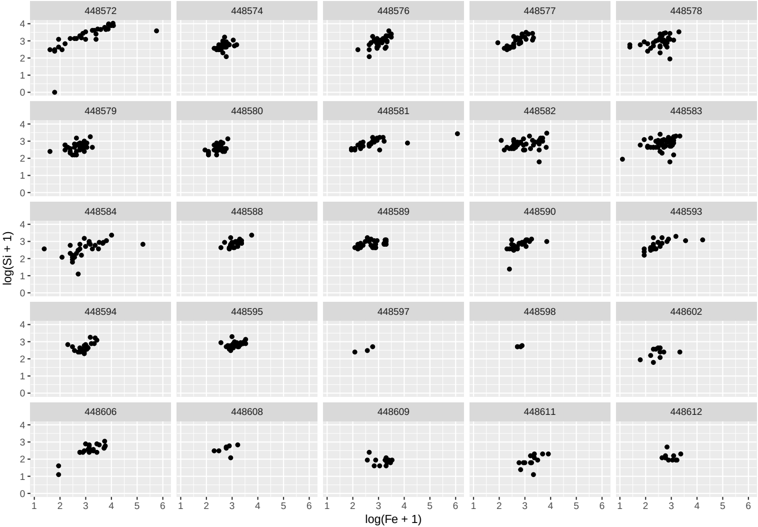


alternatively, on a log-log scale

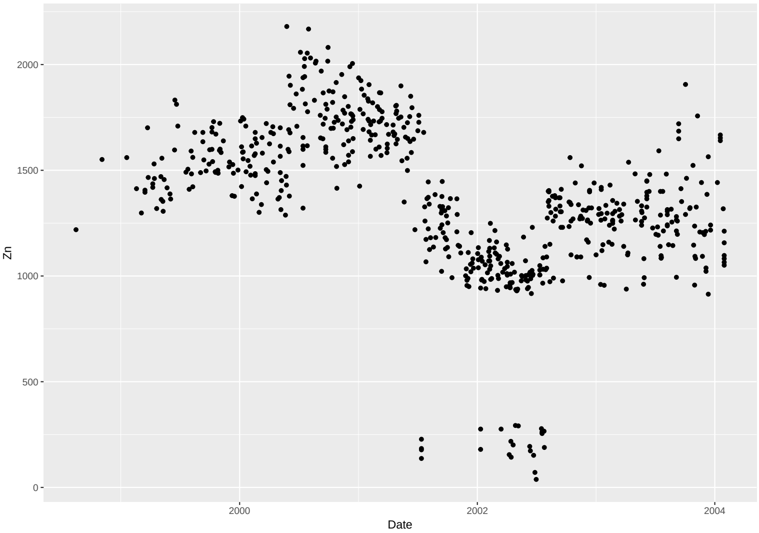


“small multiples” through faceting

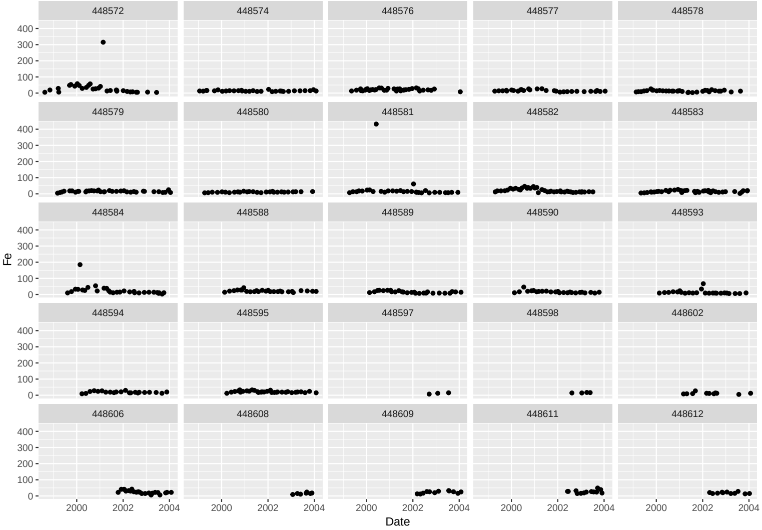
A powerful exploratory tool is to make a grid of small plots on subsets of the data.



what about that “Date” variable. . . (!)



Fe versus Date, facet by Ident



probability

the goal

Consider a variable in a dataset.

We need a mathematical model for the nature of the variation in that variable.

We need to start with a little set theory and to define what is meant by “probability.”

We will not talk philosophy (i.e. “what is the meaning of probability?”)—we will take an axiomatic approach without worrying too much about very deep meanings.

sample space

A random process will have a set of possible *outcomes* which together is called a *sample space*, often denoted by S.

e.g. toss a coin

$$S = \{H, T\}$$

e.g. roll a six-sided die

$$S = \{1, 2, 3, 4, 5, 6\}$$

e.g. toss a coin repeatedly until the first time ‘H’ appears

$$S = \{H, TH, TTH, TTTH, \ldots\}$$

e.g. run a backhoe, measuring the amount of time until it fails

$$S = (0, \infty)$$

events

An event is a subset of a sample space.

e.g. If $S = \{H, T\}$, the events are:

$$\emptyset, \{H\}, \{T\}, S$$

Note 1: H by itself is an *outcome*. The set containing $\{H\}$ is an *event*. But:

$$H \neq \{H\}$$

Note 2: the empty event is needed for technical reasons. The book tries to make a practical example which is misleading (p.39 "detecting a microscopic. . . ")

Another example: is $S = \{H, TH, TTH, TTTH, \ldots\}$ then

$$\{H, TTH, TTTTH, \ldots\}$$

is the event "odd number of tosses."

fun technicality

When the sample space is finite, or an infinite *list* of outcomes, we say the sample space is *countable*. All subsets of a countable sample space can be considered to be events.

All teaching of probability to undergraduate students is limited by some deep techicalities when the sample space is a real interval, such as with:

$$S = (0, \infty)$$

The sample space is *uncountable*, and not *all* subsets are allowed to be events.

This has something to do with the hierarchy of sizes of infinite sets. If you're interested, there is a document with the lecture materials you can read.

reminders (?!) of basic set theory

- Events are typically given names from the beginning of the capital Roman alphabet: A, B, C, A_1, A_2, A_3, $\ldots$
- The outcomes in S but *not* in A is an event called A', the *complement* of A. (aka A^c aka $\overline{A}$)
- The outcomes in both A **and** B is an event called $A \cap B$, the *intersection*.
- The outcomes in either A **or** B is an event called $A \cup B$, the *union*.
- If $A \cap B = \emptyset$ we say A and B are *disjoint* (aka *mutually exclusive*.)
- If the collection of events $\{A_1, A_2, A_3, \ldots\}$ are disjoint and

  $$\bigcup_{i=1}^{\infty} A_i = S$$

  we say the collection is a *partition* of S.
- All of the above work for finite and infinite collections

formal definition of probability

A probability is a *function*.

Given a sample space S and a collection of events $\mathcal{A}$, a *probability* is a function $P : \mathcal{A} \to \mathbb{R}$ which satisfies:

1. $P(S) = 1$
2. $P(A) \geqslant 0$
3. If $A_1, A_2, A_3, \ldots$ are disjoint then

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

"The three axioms"

One can prove that 3. also works for finite collections of disjoint events.

example of a probability function

Coin toss.

$S = \{H, T\}$

$\mathcal{A} = \Big\{\emptyset, \{H\}, \{T\}, S\Big\}$

Define P as follows:

$$\begin{aligned} P(\emptyset) &= 0 \\ P(\{H\}) &= 0.5 \\ P(\{T\}) &= 0.5 \\ P(S) &= 1 \end{aligned}$$

P satisfies the required properties.

a class of valid probability functions

If $S = \{\omega_1, \omega_2, \omega_3, \ldots\}$ is *countable* then $\mathcal{A}$ can be the collection of all subsets of S and any function P with:

- $P(\{\omega_i\}) = p_i$
- $p_i \geqslant 0$
- $\sum\limits_{i=1}^{\infty} p_i = 1$

is a probability function.

Example: toss a fair die; assign each side probability $1/6$.

Example: ("toss until first head") assign probability $(1/2)^i$ to the event "first head on i^{th} toss."

sample spaces with equally likely outcomes

Denote the size of a set A by $|A|$.

A common special case of the previous slide is with $|S| = n$ finite and $p_i = 1/n$ for all i.

In this case the probabilities will all be:

$$P(A) = \frac{|A|}{|S|}$$

This fact explains the existence of textbook sections such as our 2.3 that go on (and on and on) about counting sizes of events. Not a focus of this course.

the axiomatic approach

Some of the basic rules can be formally proven, which is great fun!

Theorem 1 $P(\emptyset) = 0$

Theorem 2 $P(A') = 1 - P(A)$

Theorem 3 If $A \subset B$ then $P(A) \leqslant P(B)$

Theorem 4 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Corollaries to Theorem 4: $P(A \cup B) = P(A) + P(B)$ when A and B are disjoint, and $P(A \cup B) \leqslant P(A) + P(B)$ (always).

the axiomatic approach

Some of the basic rules can be formally proven, which is great fun!

Theorem 1 $P(\emptyset) = 0$

Theorem 2 $P(A') = 1 - P(A)$

Theorem 3 If $A \subset B$ then $P(A) \leqslant P(B)$

Theorem 4 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Corollaries to Theorem 4: $P(A \cup B) = P(A) + P(B)$ when A and B are disjoint, and $P(A \cup B) \leqslant P(A) + P(B)$ (always).

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

Let's use a "personal probability" philosophy for the momemnt.

What if I peek and tell you "Actually, B occurred". What is the (*your?*) probability 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 people use a "sample space restriction" approach in these simple cases.

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.

Fun fact: For a fixed B with $P(B) > 0$, the function $P_B(A) = P(A|B)$ is a probability function. (You can prove this.)

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 $P(B) = 0.00212$.

The probability of a randomly selected Canadian being HIV positive and testing positive is:

$$P(A \cap B) = P(A|B)P(B) = 0.0021094$$

useful expressions for calculation - II

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

$$\begin{aligned} P(A) &= P\left(\bigcup_i (A \cap B_i)\right) \\ &= \sum_i P(A \cap B_i) \\ &= \sum_i P(A|B_i)P(B_i) \end{aligned}$$

Common simple version: $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$

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

useful expressions for calculation - III

We can now calculate $P(A)$, the probability of a randomly selected Canadian testing positive.

$$\begin{aligned}P(A) &= P(A|B)P(B) + P(A|B^c)P(B^c)\\ &= 0.995 \cdot 0.00212 + 0.005 \cdot (1 - 0.00212)\\ &= 0.0070988\end{aligned}$$

The simple formula gets a grandiose title: ***"THE! LAW! OF! TOTAL! PROBABILITY!!!!!"***

Now, in the HIV example, we also might be interested in $P(B|A)$, the probability that someone is HIV+ given that they test positive.

$P(B|A)$

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 in general

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

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

independence

motivation - revisit the die toss example

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

$$A = \{2, 5\},$$
$$B = \{2, 4, 6\}$$

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:

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

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.

$A_1, A_2, A_3, \ldots$ (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:

$$\mathcal{A} = \{\emptyset, A, A^c, S\}$$
$$\mathcal{B} = \{\emptyset, B, B^c, S\}$$

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 that eventually we will pretty much forget even exist

sample spaces: too general?

The goal for *our* study of probability is to define a mathematical model for a variable in a dataset.

Sample spaces are needed to define the important building block P (probability function), but are not usually convenient.

e.g. six horses in a race, labelled 'A' through 'F'. You bet \$2 on 'D' to win. You will get \$6.75 if 'D' wins.

What is the sample space? Some options:

- all 6! possible orders of finish: $\{EDBFCA, ECADBF, BDAECF, BEDCFA\ldots\}$.
- all possible combinations of the six horse distances run as functions of time.

But really all you care about is whether you suffer a \$2 loss or a \$4.75 profit, and it turns out to be better just to focus on that.

a new style of function to add to the stable

We will now consider functions that have a sample space S as a domain and range $\mathbb{R}$, e.g.

Element of S	Value in $\mathbb{R}$
EDBFCA	-2.00
ECADBF	-2.00
BDAECF	-2.00
BEDCFA	-2.00
$\vdots$	$\vdots$

Note the impossibility of doing things like drawing pictures of such functions.

So if all the usual function things like derivatives, local maxima, integrals etc. are not of interest, what do we do with these things?

All we care about are the *distributions* of such functions—roughly speaking the possible values in $\mathbb{R}$ and their probabilities.

“random variables”

Sadly these functions have a terrible name.

Definition: A *random variable* is a real-valued function of a sample space.

More examples:

Toss to first head: Count the number of tosses of a coin until the first `H` appears.

Element of S	Value in $\mathbb{R}$
H	1
TH	2
TTH	3
TTTH	4
TTTTH	5
$\vdots$	$\vdots$

examples of random variables

See if a product is defective: Select an item at random from a factory. See if it is defective. Define the following function:

Element of S	Value in $\mathbb{R}$
Defective	1
Not Defective	0

Seems like a stupid example, but this is merely an instance of one of the most important random variables of all. To generalize:

"Bernoulli trial" (book: *Bernoulli random variable*) Observe a random process to see if an event A occurred. The *indicator function* I_A which takes on the value 1 if A occurred and 0 otherwise is called a Bernoulli trial.

more examples of random variables

Bus stop Busses arrive at a stop at 10 minute intervals. You arrive at the bus stop at a random time. Observe the amount of time you have to wait for the next bus.

This random variable is of a fundamentally different character from the other examples. This random variable could take on *any* real number between 0 and 10 (using minutes as the unit of measure.)

Failure time(s) Observe a pump until its bearing fails. Observe a pump until its seal fails. Observe a pump until either component fails.

These random variables could take on any positive real number.

notation and naming conventions

Random variables are given names that tend to be capital Roman letters near the end of the alphabet.

$$X : S \to \mathbb{R}$$

Usual names: X, Y, Z, X_1, X_2, X_3, etc.

(Just like in calculus: f, g, h, f_1, f_2, etc.)

In calculus one tends to conflate f (the function name) and $f(x)$ (the value at a generic x)

In probability we just use the function name, i.e. X.

Is notation important? If you like to watch the world burn put this on a calculus exam:

$$\int_{0}^{2\pi} sin(f)\, df$$

values of random variables imply events

For any random variable X, any subset of the real line you could think of implies an event.

Examples:

Toss to first head: Let X be the number of tosses.

X taking on any of these values	Implies this event
$X \in \{1\}$	$\{H\}$
$X \in [2, 4]$	$\{TH, TTH, TTTH\}$
$X \in [2, 4)$	$\{TH, TTH\}$
$X \in (-9, -3.2]$	$\emptyset$
$X \in [0.5, 1.3) \cup (\pi, \pi + 1)$	$\{H, TTTH\}$
$X \in [0, \infty)$	S

distribution of a random variable

The distribution of a random variable is the mapping between values of X and their probabilities of the implied events.

(Over-)simplified: the values, and their probabilities.

Examples:

X taking on any of these values	Probability
$X \in \{1\}$	$\frac{1}{2}$
$X \in [2, 4]$	$\frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{7}{16}$
$X \in [2, 4)$	$\frac{1}{4} + \frac{1}{8} = \frac{3}{8}$
$X \in (-9, -3.2]$	0
$X \in [0.5, 1.3) \cup (\pi, \pi + 1)$	$\frac{1}{2} + \frac{1}{16} = \frac{9}{16}$
$X \in [0, \infty)$	1

distribution in the bus stop example

X is the amount of time you wait for the bus, with the idea that the bus could come at any random time in the next 10 minutes "uniformly"

X taking on any of these values	Probability
$X \in [2, 4]$	$\frac{2}{10}$
$X \in [3, 5]$	$\frac{2}{10}$
$X \in (-9, -3.2]$	0
$X \in [0.5, 1.3) \cup (\pi, \pi + 1)$	$\frac{0.8}{10} + \frac{\pi}{10}$
$X \in [2, 2.1]$	0.01
$X \in [2, 2.001]$	0.0001
$X \in \{2\}$	0

more notation

Here's how we will actually write the probability statements:

Inconvenient	Usual Notation
$X \in [2, 4]$	$P(2 \leqslant X \leqslant 4) = \frac{2}{10}$
$X \in [3, 5]$	$P(3 \leqslant X \leqslant 5) = \frac{2}{10}$
$X \in (-9, -3.2]$	$P(-9 < X \leqslant -3.2) = 0$
$X \in \{2\}$	$P(X = 2) = 0$

distributions and their representations

When you know the distribution of a random variable, you know everything *(as far as probability theory goes.)*

It's clear from the examples that "distribution" is a complex object, so we'll need convenient representations for them. Here is the first one.

("Theorem":)the distribution of any random variable X is completely determined by the following function $F : \mathbb{R} \to \mathbb{R}$, defined at every $x \in \mathbb{R}$:

$$F(x) = P(X \leqslant x).$$

F is called the *cumulative distribution function* of X, or cdf for short.

cumulative distribution functions

All random variables have a cumulative distribution function, defined as follows for an r.v. named X:

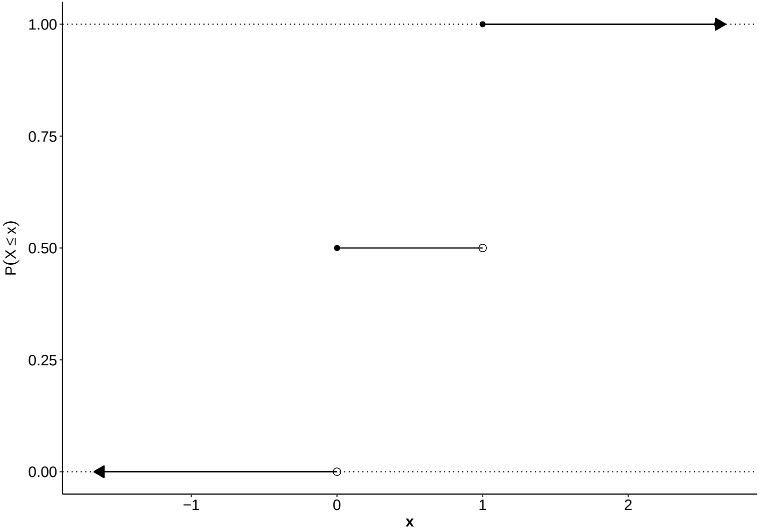
$$F(x) = P(X \leqslant x)$$

You could make a picture of a cdf.

Example: Toss a coin; map `H` to 1 and `T` to 0, calling the map X. Then:

$$F(x) = \begin{cases} 0 & : x < 0 \\ 0.5 & : 0 \leqslant x < 1 \\ 1 & : x \geqslant 1 \end{cases}$$

picture of cdf



defining properties of a cdf

A function is a cdf if and only if:

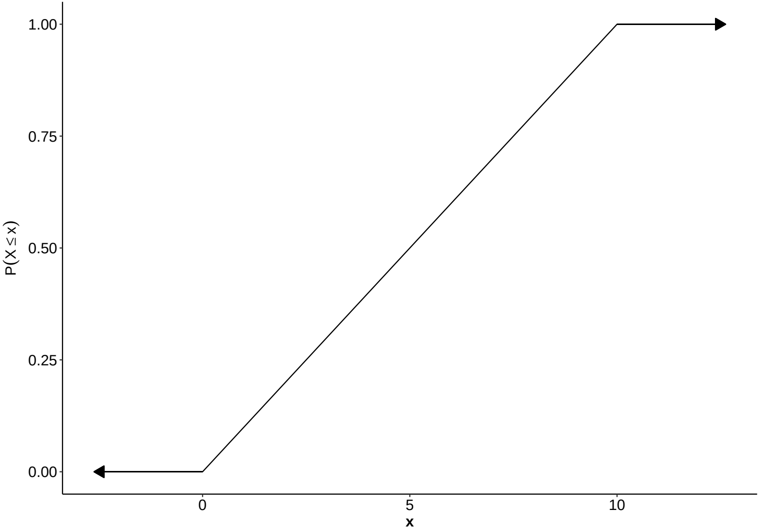
1. $\lim_{x \to -\infty} F(x) = 0$ (“Starts at 0”)
2. $\lim_{x \to +\infty} F(x) = 1$ (“Ends at 1”)
3. it is non-decreasing, i.e. $x \leqslant y$ implies $F(x) \leqslant F(y)$
4. it is “right-continuous” (technical definition: $\lim_{x \to x_0^+} F(x) = F(x_0)$)

Example: waiting for the bus for between 0 and 10 minutes “uniformly”.

$$F(x) = \begin{cases} 0 & : x < 0 \\ x/10 & : 0 \leqslant x < 10 \\ 1 & : x \geqslant 10 \end{cases}$$

Equality vs. inequality not really important in a “continuous” probability model.

picture of this cdf



a taxonomy of random variables

It is possible to classify random variables:

1. *discrete* random variables take on a finite or countably infinite number of outcomes.
    - its cdf will be a *step function*
2. *continuous* random variables take on values in intervals
    - its cdf will be a *continuous function*
3. Neither discrete nor continuous.

As an example of the latter, consider the time-to-failure of an electronic component that:

- fails immediately the first time you try it, with probability 0.01
- works immediately with probability 0.99 and subsequently fails according to some continuous probability model TBA.

discrete random variables

CDFs are nice because *all* random variables have one, but they aren't the most natural representations of distributions.

For a discrete random variable X, the function that maps the individual outcomes to their probabilities is more natural.

Example: Let X be the number of tosses of a coin until the first `H` appears. This function maps individual outcomes to probabilities:

$$p(x) = f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x \in \{1, 2, 3, \ldots\}$$

Such a function is (best) called a *probability mass function* or pmf. I tend to use $p(x)$ notation for pmfs.

Textbook notes: the book uses $f(x)$ notation, which I dislike. It also gives the following (terrible) synonyms:

- "probability function" (name already taken by P!)
- "probability distribution" (name already being used for a fundamental concept!)

more pmf examples

See if a product is defective: A factory makes a defective item with probability p. Select an item at random from a factory. Let $X = 1$ if the item is defective, and let $X = 0$ otherwise.

The pmf of X is:

$$p(x) = P(X = x) = \begin{cases} p & : x = 1 \\ 1 - p & : x = 0 \end{cases}$$

More compact version: $p(x) = p^x(1 - p)^{1-x}$

defining properties of pmf

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

1. $$p(x) \geqslant 0$$

2. $$\sum_{\{x \,|\, P(X=x)>0\}} p(x) = 1$$

checking if a function is a valid pmf

I said this function is a pmf. Is it?

$$p(x) = f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x \in \{1, 2, 3, \ldots\}$$

Verify:

1. $p(x) \geqslant 0$
2. Fact: $\sum\limits_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ for $0 < r < 1$. So:

$$\sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \sum_{x=0}^{\infty} \frac{1}{2}\left(\frac{1}{2}\right)^x = 1$$

a pmf completely characterizes a discrete distribution

I told you a cdf completely characterizes any distribution, which is a fact you'll have to take on buffy.

A discrete random variable has a pmf. Does the pmf characterize the distribtuion?

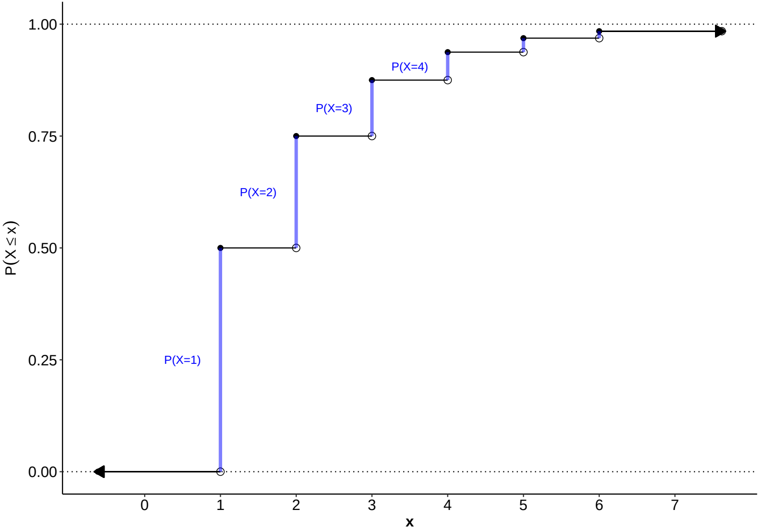
Yes, because you can compute a cdf from a pdf and vice versa. "Obviously:"

$$F(x) = \sum_{y \leqslant x} p(y)$$

For the reverse direction you take the jump points of the cdf and determine the magnitude of the jump.

possibly easier to see than to understand the formal statement

The cdf of $X =$ "toss to first H", with pmf values in blue:



cumulative distribution functions

All random variables have a cumulative distribution function, defined as follows for an r.v. named X:

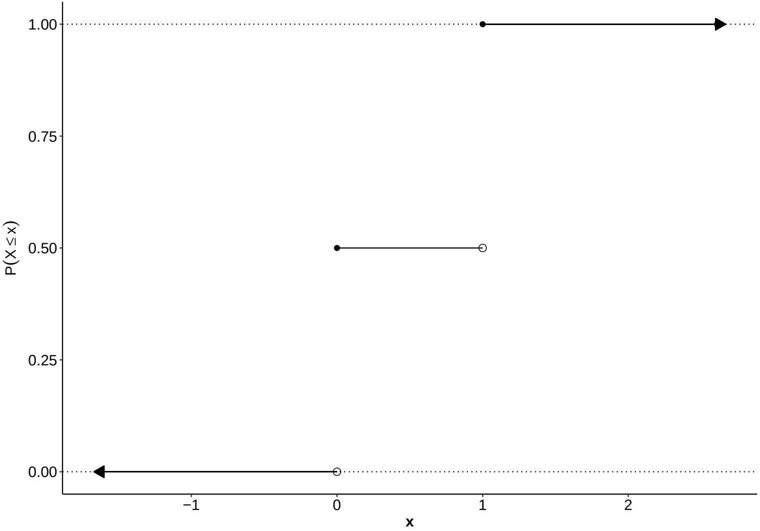
$$F(x) = P(X \leqslant x)$$

You could make a picture of a cdf.

Example: Toss a coin; map `H` to 1 and `T` to 0, calling the map X. Then:

$$F(x) = \begin{cases} 0 & : x < 0 \\ 0.5 & : 0 \leqslant x < 1 \\ 1 & : x \geqslant 1 \end{cases}$$

picture of cdf



defining properties of a cdf

A function is a cdf if and only if:

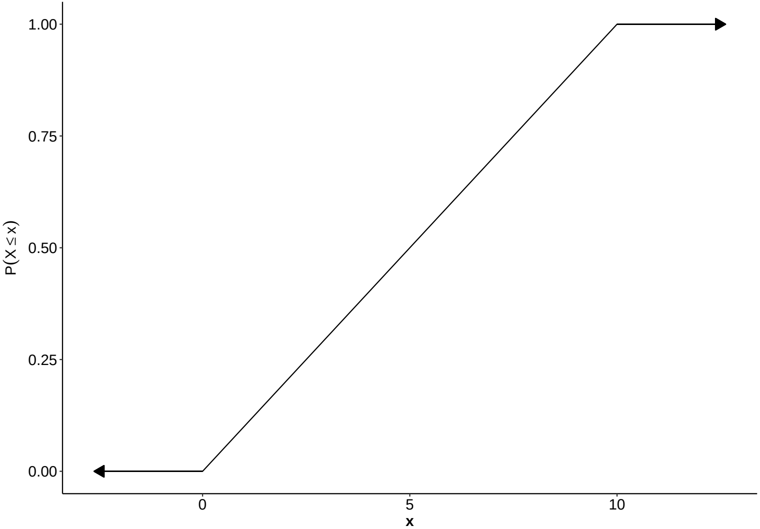
1. $\lim_{x \to -\infty} F(x) = 0$ ("Starts at 0")
2. $\lim_{x \to +\infty} F(x) = 1$ ("Ends at 1")
3. it is non-decreasing, i.e. $x \leqslant y$ implies $F(x) \leqslant F(y)$
4. it is "right-continuous" (technical definition: $\lim_{x \to x_0^+} F(x) = F(x_0)$)

Example: waiting for the bus for between 0 and 10 minutes "uniformly".

$$F(x) = \begin{cases} 0 & : x < 0 \\ x/10 & : 0 \leqslant x < 10 \\ 1 & : x \geqslant 10 \end{cases}$$

Equality vs. inequality not really important in a "continuous" probability model.

picture of this cdf



a taxonomy of random variables

It is possible to classify random variables:

1. *discrete* random variables take on a finite or countably infinite number of outcomes.
    - its cdf will be a *step function*
2. *continuous* random variables take on values in intervals
    - its cdf will be a *continuous function*
3. Neither discrete nor continuous.

As an example of the latter, consider the time-to-failure of an electronic component that:

- fails immediately the first time you try it, with probability 0.01
- works immediately with probability 0.99 and subsequently fails according to some continuous probability model TBA.

discrete random variables

CDFs are nice because *all* random variables have one, but they aren't the most natural representations of distributions.

For a discrete random variable X, the function that maps the individual outcomes to their probabilities is more natural.

Example: Let X be the number of tosses of a coin until the first `H` appears. This function maps individual outcomes to probabilities:

$$p(x) = f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x \in \{1, 2, 3, \ldots\}$$

Such a function is (best) called a *probability mass function* or pmf. I tend to use $p(x)$ notation for pmfs.

Textbook notes: the book uses $f(x)$ notation, which I dislike. It also gives the following (terrible) synonyms:

- "probability function" (name already taken by P!)
- "probability distribution" (name already being used for a fundamental concept!)

more pmf examples

See if a product is defective: A factory makes a defective item with probability p. Select an item at random from a factory. Let $X = 1$ if the item is defective, and let $X = 0$ otherwise.

The pmf of X is:

$$p(x) = P(X = x) = \begin{cases} p & : x = 1 \\ 1 - p & : x = 0 \end{cases}$$

More compact version: $p(x) = p^x (1 - p)^{1-x}$

defining properties of pmf

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

1. $$p(x) \geqslant 0$$

2. $$\sum_{\{x \,|\, P(X=x)>0\}} p(x) = 1$$

checking if a function is a valid pmf

I said this function is a pmf. Is it?

$$p(x) = f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x \in \{1, 2, 3, \ldots\}$$

Verify:

1. $p(x) \geqslant 0$
2. Fact: $\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ for $0 < r < 1$. So:

$$\sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \sum_{x=0}^{\infty} \frac{1}{2} \left(\frac{1}{2}\right)^x = 1$$

a pmf completely characterizes a discrete distribution

I told you a cdf completely characterizes any distribution, which is a fact you'll have to take on buffy.

A discrete random variable has a pmf. Does the pmf characterize the distribtuion?

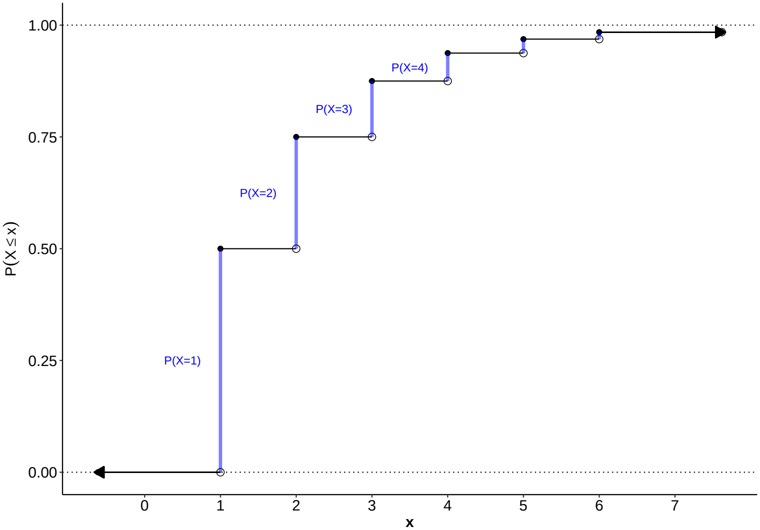
Yes, because you can compute a cdf from a pdf and vice versa. "Obviously:"

$$F(x) = \sum_{y \leqslant x} p(y)$$

For the reverse direction you take the jump points of the cdf and determine the magnitude of the jump.

possibly easier to see than to understand the formal statement

The cdf of $X =$ "toss to first H", with pmf values in blue:



cumulative distribution functions

Recall the cdf:

$$F(x) = P(X \leqslant x),$$

which completely decribes the distribution of X.

discrete random variables

CDFs are nice because *all* random variables have one, but they aren't the most natural representations of distributions.

For a discrete random variable X, the function that maps the *individual possible values* to their probabilities is more natural.

Example: Let X be the number of tosses of a coin until the first `H` appears. This function maps individual possible values to probabilities:

$$p(x) = f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x \in \{1, 2, 3, \ldots\}$$

Such a function is (best) called a *probability mass function* or pmf. I tend to use $p(x)$ notation for pmfs.

Textbook notes: the book uses $f(x)$ notation, which I dislike. It also gives the following (terrible) synonyms:

- "probability function" (name already taken by P!)
- "probability distribution" (name already being used for a fundamental concept!)

more pmf examples

See if a product is defective: A factory makes a defective item with probability p. Select an item at random from a factory. Let $X = 1$ if the item is defective, and let $X = 0$ otherwise.

The pmf of X is:

$$p(x) = P(X = x) = \begin{cases} p & : x = 1 \\ 1 - p & : x = 0 \end{cases}$$

More compact version: $p(x) = p^x(1-p)^{1-x}$

defining properties of pmf

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

1. $$p(x) \geqslant 0$$

2. $$\sum_{\{x \mid P(X=x)>0\}} p(x) = 1$$

checking if a function is a valid pmf

I said this function is a pmf. Is it?

$$p(x) = f(x) = P(X = x) = \left(\frac{1}{2}\right)^x, \quad x \in \{1, 2, 3, \ldots\}$$

Verify:

1. $p(x) \geqslant 0$
2. Fact: $\sum\limits_{x=0}^{\infty} ar^x = \frac{a}{1-r}$ for $0 < r < 1$. So:

$$\sum_{x=1}^{\infty} \left(\frac{1}{2}\right)^x = \sum_{x=0}^{\infty} \frac{1}{2} \left(\frac{1}{2}\right)^x = 1$$

a pmf completely characterizes a discrete distribution

I told you a cdf completely characterizes any distribution, which is a fact you'll have to take on buffy.

A discrete random variable has a pmf. Does the pmf characterize the distribtuion?

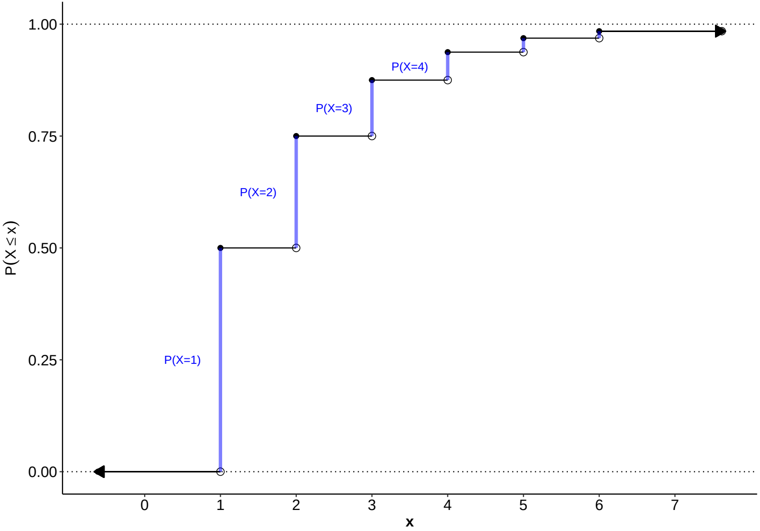
Yes, because you can compute a cdf from a pdf and vice versa. "Obviously:"

$$F(x) = \sum_{y \leqslant x} p(y)$$

For the reverse direction you take the jump points of the cdf and determine the magnitude of the jump.

possibly easier to see than to understand the formal statement

The cdf of $X =$ “toss to first H”, with pmf values in blue:



enormous cat

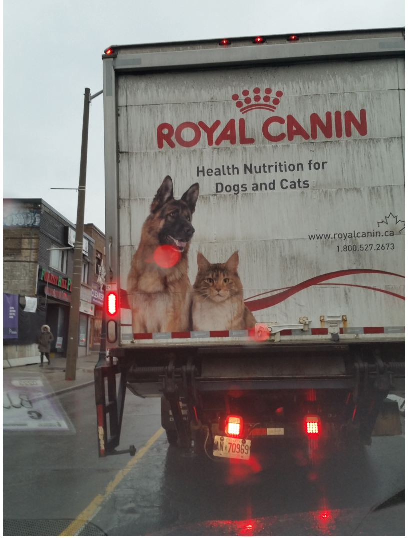


Figure 3: alt text

continuous random variables

For a random process taking on values in real intervals, we saw it made sense for $P(X = x) = 0$ for any particular value of x (e.g. bus stop example.)

That condition could be taken as a definition of "continuous random variable".

We're mainly concerned with probabilities like $P(a < X \leqslant b)$, which could be calculated using $F(b) - F(a)$, but there's another way.

If there is a ("Riemann integrable") function f such that:

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

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 ∞.

example - bus stop

The bus comes every 10 minutes and you arrive at random “uniformly”. X is the waiting time for the bus.

Let:

$$f(x) = \begin{cases} \frac{1}{10} & : 0 < x < 10 \\ 0 & : \text{otherwise} \end{cases}$$

This density gives us all the probabilities such as:

$$P(2 < X \leqslant 4) = \int_{2}^{4} \frac{1}{10}\, dx = \frac{2}{10} \qquad P(X = 2) = \int_{2}^{2} \frac{1}{10}\, dx = 0$$

a density completely characterizes a distribution

Since:

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

one gets $F'(x) = f(x)$, so F and f contain equivalent information.

Defining characteristics: A function f is a density as long as $f \geqslant 0$ and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

another density example

Consider:

$$f(x) = \begin{cases} e^{-x} & : x > 0 \\ 0 & : x \leqslant 0 \end{cases}$$

It satisfies the requirements to be a density, since $f \geqslant 0$ and:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} e^{-x}\,dx = \left[-e^{-x}\right]_0^{\infty} = 1$$

Suppose X has this density. Calculate $P(X > 1)$ and determine the cdf of X...

density - meaning and interpretation

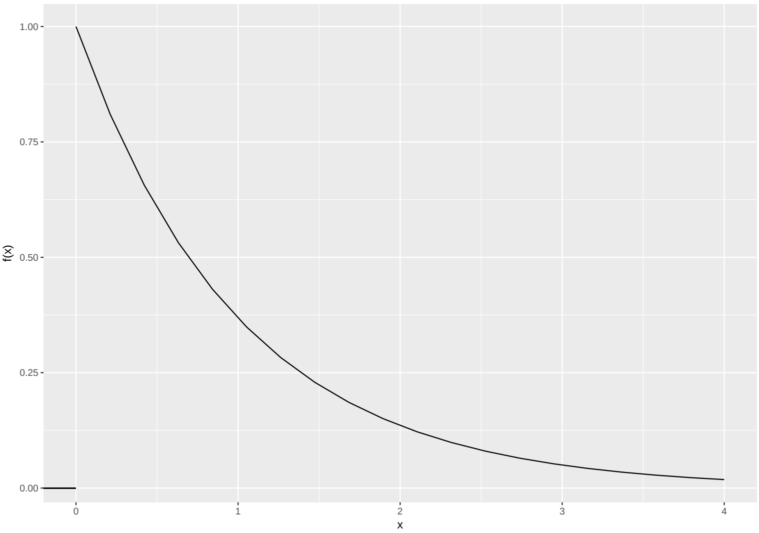
Advice: *Always* think of a density as living inside its integral.

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

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

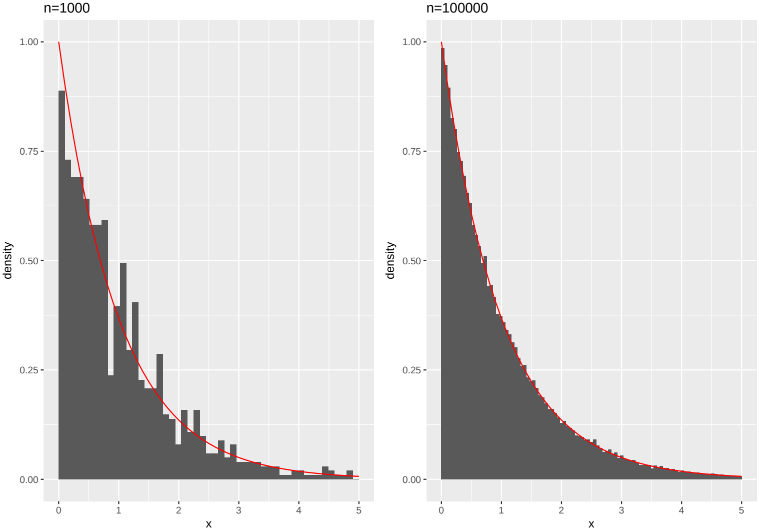
Pictures of densities can be useful, to show relative differences in probabilities.

illustration using e^{-x} density



histogram as “density estimator”

A density can be thought of as the “limit of histograms”.



a note on “identically distributed”

The distribution is all we care about.

So if X_1 and X_2 have the same distributions, they are effectively the same (even if they are not the same functions.)

For example, roll a fair die so that $S = \{1, 2, 3, 4, 5, 6\}$.

Define:

$$X_1 = \begin{cases} 1 & : \text{3 or 4 appears} \\ 0 & : \text{otherwise} \end{cases} \qquad \text{and} \qquad X_2 = \begin{cases} 1 & : \text{5 or 6 appears} \\ 0 & : \text{otherwise} \end{cases}$$

X_1 and X_2 are not the same functions. But the have the same p.m.f.:

$$p(x) = \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{1-x}, \ x \in \{0, 1\}.$$

We say X_1 and X_2 are *identically distributed*.

joint distributions

more than one random variable at a time

A dataset will usually have more than one variable. They might be modeled at the same time.

Certainly a dataset will have more than one observation, so considering multiple random variables is essential.

The approach is to consider a random vector such as (X, Y), (X_1, X_2), $(X_1, X_2, \ldots X_n)$ and so on. At first we'll consider two at a time: (X, Y).

As usual we are interested in the *distribution* of, say, (X, Y), which now means the collection of things like $P((X, Y) \in A)$ for $A \in \mathbb{R}^2$.

Main interest is in statements like:

$$P(X = x, Y = y) \qquad \text{or} \qquad P(a < X < b, c < Y < d)$$

where the “comma” notation is a compact way of writing, say:

$$P(\{X = x\} \cap \{Y = y\})$$

cdf, pmf, pdf

The natural ways to characterize joint distributions are still probability mass functions and probability density functions.

They also have cdfs:

$$F(x, y) = P(X \leqslant x, Y \leqslant y)$$

but we won't focus much on these.

a joint pmf

A gas distribution company has pipes with diameters 1, 1.5, and 1.75 inches. The pipes are used at pressures of 2, 1, and 0.5 pounds per square inch.

Pick a pipe at random and denote its diameter by X and its pressure by Y.

X and Y might have the following *joint probability mass function*:

		X	
Y	1	1.5	1.75
0.5	0.075	0.100	0.150
1	0.110	0.080	0.090
2	0.160	0.140	0.095

e.g. the probability that a randomly selected pipe has diameter $X = 1.5$ and pressure $Y = 0.5$ is:

$$P(X = 1.5, Y = 0.5) = 0.1$$

joint pmf properties; joint density

A joint pmf is still just a pmf. It must be non-negative, and its positive values must add up to 1.

(X, Y) are (jointly) continuous if they have a joint density f such that:

$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x, y)\, dxdy$$

A joint density also must be non-negative and integrate to 1 over all $\mathbb{R}^2$.

joint density example - I

Two electronic components fail at times X and Y according to the following joint density (measured in years):

$$f(x,y) \begin{cases} 2e^{-x-2y} & : x > 0, y > 0 \\ 0 & : \text{otherwise} \end{cases}$$

Is this actually a valid joint density? It is non-negative, and:

$$\iint\limits_{\mathbb{R}^2} f(x,y)\,dxdy = \int\limits_0^\infty \int\limits_0^\infty 2e^{-x-2y}\,dxdy = \int\limits_0^\infty e^{-x}\,dx \int\limits_0^\infty 2e^{-2y}\,dy = 1$$

joint density example - II

What is the probability that the first component fails before the second? In other words, what is $P(X < Y)$?

Answer: integrate the joint density over the region where $x < y$.

$$\begin{aligned}
\int\limits_{0}^{\infty}\int\limits_{0}^{y} 2e^{-x-2y}\,dx\,dy &= \int\limits_{0}^{\infty}\left[-2e^{-x-2y}\right]_{x=0}^{y}\,dy \\
&= \int\limits_{0}^{\infty} 2e^{-2y} - 2e^{-3y}\,dy \\
&= \left[-e^{-2y} + \frac{2}{3}e^{-3y}\right]_{y=0}^{\infty} \\
&= \frac{1}{3}
\end{aligned}$$

marginal distributions

A joint distribution contains *all* information about both X and Y together.

"*All*" includes information about X alone, and about Y alone. And it's easy to get this information from the joint distribution.

For example, look again at the joint distribution for diameter X and pressure Y of the randomly selected pipe:

	X		
Y	1	1.5	1.75
0.5	0.075	0.100	0.150
1	0.110	0.080	0.090
2	0.160	0.140	0.095

A statement about X alone could be something like $P(X = 1)$, which is just $0.075 + 0.11 + 0.16 = 0.345$

marginal pmf

Given the joint pmf $p(x, y)$ for X and Y, the marginal pmf for X is:

$$p_X(x) = \sum_y p(x, y)$$

Here are the "marginals" for both X and Y in the gas example:

	X			
Y	1	1.5	1.75	Marginal
0.5	0.075	0.100	0.150	0.325
1	0.110	0.080	0.090	0.280
2	0.160	0.140	0.095	0.395
Marginal	0.345	0.320	0.335	1.000

marginal density

Reconsider the two electronic components example:

$$f(x,y) \begin{cases} 2e^{-x-2y} & : x > 0, y > 0 \\ 0 & : \text{otherwise} \end{cases}$$

A statement about, say, Y alone might be $P(Y > 1)$, which would be the integral on the entire half plane $y > 1$:

$$\iint\limits_{y>1} f(x,y)\,dxdy = \int\limits_1^\infty \int\limits_0^\infty 2e^{-x-2y}\,dxdy = \int\limits_1^\infty 2e^{-2y}\,dy = e^{-2}$$

But in the end, the calculation only ever involved $2e^{-2y}$ once x was "integrated out."

marginal density

Given the joint density $f(x, y)$ for X and Y, the marginal density for X is:

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

Marginal densities can't really be visualized. At best they are a "projection" onto one or the other axis.

e.g. Consider the following function:

$$f(x, y) = \begin{cases} 2 & : 0 < x < 1,\ 0 < y < 1,\ x + y < 1 \\ 0 & : \text{otherwise} \end{cases}$$

This is a valid density. Let X and Y have this joint density.

marginal density

The marginal density for X will be 0 outside $x \in (0,1)$. Otherwise:

$$\begin{aligned} f_X(x) &= \int\limits_{-\infty}^{\infty} f(x,y)\,dy \\ &= \int\limits_{0}^{1-x} 2\,dy \\ &= 2(1-x) \end{aligned}$$

conditional distributions

Consider again the gas pipe joint (and marginal) distributions:

	X			
Y	1	1.5	1.75	Marginal
0.5	0.075	0.100	0.150	0.325
1	0.110	0.080	0.090	0.280
2	0.160	0.140	0.095	0.395
Marginal	0.345	0.320	0.335	1.000

What are the probabilities of X taking on any of its three possible values *given* $Y = 2$, *say.*

$$P(X = 1|Y = 2) = \frac{P(X = 1, Y = 2)}{P(Y = 2)} = \frac{0.16}{0.395}$$

$$P(X = 1.5|Y = 2) = \frac{0.14}{0.395} \qquad P(X = 1.75|Y = 2) = \frac{0.095}{0.395}$$

conditional pmf and conditional density

Given the joint pmf $p(x, y)$ for X and Y, the conditional pmf for X given $Y = y$ is:

$$p(x|y) = \frac{p(x, y)}{p_Y(y)},$$

provided $P(Y = y) > 0$.

Given the joint density $f(x, y)$ for X and Y, the conditional density for X given $Y = y$ is:

$$f(x|y) = \frac{f(x, y)}{f_Y(y)},$$

provided $f_Y(y) > 0$.

conditional density example “uniform on a triangle”

Consider again X and Y with the following density.

$$f(x,y) = \begin{cases} 2 & : 0 < x < 1,\ 0 < y < 1,\ x + y < 1 \\ 0 & : \text{otherwise} \end{cases}$$

The marginal density for X was found to be $f_X(x) = 2(1 - x)$ for $0 < x < 1$.

The conditional density of Y given $X = 0.8$ will be 0 when y is outside $(0, 0.2)$. Otherwise it will be:

$$f(y|x = 0.8) = \frac{2}{2 \cdot (1 - 0.8)} = 5$$

Unlike marginal densities, conditional densities can sometimes be visualized. They are just “slices” of the joint density (normalized to 1.)

conditional density example

Reconsider the two electronic components example:

$$f(x, y) \begin{cases} 2e^{-x-2y} & : x > 0, y > 0 \\ 0 & : \text{otherwise} \end{cases}$$

The marginal density for Y is $2e^{-2y}$ on $y > 0$.

The conditional density for X given $Y = 2$ is:

$$e^{-x} \text{ on } x > 0$$

Heck, the conditional density for X given Y equals *anything* greater than 0 is still always e^{-x} on $x > 0$.

That's because X and Y have a special relationship. . . to be revisited.

independence

In the "two electronic component" example, the conditional density for X given $Y = y$ (no matter what $y > 0$) never changes:

$$f(x|y) = f(x)$$

Knowledge of the outcome of Y doesn't tell you anything about the distribution of X.

In the "uniform on a triangle" example, the conditional density for Y given $X = x$ is always going to be (for $0 < x < 1$):

$$f(y|x) = \frac{2}{2(1-x)} = \frac{1}{1-x}$$

for $0 < y < x$ and 0 otherwise. So knowledge of the outcome of X tells something about the distribution of Y.

independence

In the gas pipes pmf:

	X			
Y	1	1.5	1.75	Marginal
0.5	0.075	0.100	0.150	0.325
1	0.110	0.080	0.090	0.280
2	0.160	0.140	0.095	0.395
Marginal	0.345	0.320	0.335	1.000

the conditional distributions for X given $Y = y$ are all different (as well as those for Y given $X = x$.)

Definition: X and Y are *independent* (or $X \perp Y$) if:

$$p(x, y) = p_X(x)p_Y(y) \qquad \text{(discrete)} \qquad\qquad f(x, y) = f_X(x)f_Y(y) \qquad \text{(continuous)}$$

gas pipe diameters and pressures “made” independent

Suppose the marginal distributions are the same. What would the joint pmf have to be?

	X			
Y	1	1.5	1.75	Marginal
0.5	$0.345 \cdot 0.325$			0.325
1			$0.335 \cdot 0.280$	0.280
2		$0.320 \cdot 0.395$		0.395
Marginal	0.345	0.320	0.335	1.000

relationship with independence of events

For an event A from a sample space S, define the *indicator random variable*:

$$I_A(\omega) = \begin{cases} 1 & : \omega \in A \\ 0 & : \omega \notin A \end{cases}$$

This is the general "Bernoulli trial" and one of the most important random variables there is.

relationship with independence of events

Suppose A and B are events in S. Consider the joint pmf for A and B:

	I_A		
I_B	0	1	Marginal
0	$P(A' \cap B')$	$P(A \cap B')$	$P(B')$
1	$P(A' \cap B)$	$P(A \cap B)$	$P(B)$
Marginal	$P(A')$	$P(A)$	

Now suppose $A \perp B$. Recall that this is actually a “strong” statement equivalent to $A \perp B'$ and $A' \perp B'$ and $A' \perp B$.

So it is also equivalent to the independence of the random variables I_A and I_B.

The formal definition of independence is:

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$$

for any $C, D \subset \mathbb{R}$.

notes on verifying independence

First of all, in practice independence continues to be someone one generally *assumes*.

But one still needs to know how to verify independence, which is easy for continuous X and Y:

- the joint density factors:
$$f(x,y) = g(x)h(y)$$
- and the non-zero region of f is a "rectangle"

Examples:

- $f(x,y) = x + y$ on $0 < x, y < 1$ and 0 otherwise.
- $f(x,y) = 24xy$ on $x > 0, y > 0, 0 < x + y < 1$ and 0 otherwise.

extension to more than two random variables - a few illustrations

The concepts of joint and marginal distributions extend to n random variables at a time. Things get ugly fast. Here is a taste....

Joint pmf of X_1, X_2, X_3:

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

Marginal for X_2:

$$p(x_2) = \sum_{x_1} \sum_{x_3} p(x_1, x_2, x_3)$$

Joint density for Y_1, Y_2, Y_3:

$$P(a_1 < Y_1 < b_1, a_2 < Y_2 < b_2, a_3 < Y_3 < b_3) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} f(y_1, y_2, y_3)\, dy_3\, dy_2\, dy_1$$

extension to more than two random variables - a few illustrations

The marginal density for X_3:

$$f_{X_3}(x_3) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3)\, dx_1\, dx_2$$

Important: Random variables $X_1, X_2, \ldots, X_n$ are *independent* if:

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

(continuous version with densities. . . discrete version is similar.)

This is conceptually important as the basis for the mathematical model of the observations in a dataset.

the expected value operator

the mean of a distribution

Recall the sample average:

$$\overline{x} = \sum_{i=1}^{n} x_i \cdot \frac{1}{n}$$

which can be considered as a weighted sum with weights $w_i = 1/n$.

A (discrete) random variable can have an “average”, which is a weighted sum of the outcomes with their probabilities as weights.

BIG MONEY. We play a gambling game called BIG MONEY. Roll a die. This is your outcome after each play of BIG MONEY:

Roll	1, 2	3	4, 5	6
Outcome $	-2	0	1	2
Probability	2/6	1/6	2/6	1/6

BIG MONEY is a “fair game”

Your *expected* financial outcome is (theoretically):

$$-2\frac{2}{6} + 0\frac{1}{6} + 1\frac{2}{6} + 2\frac{1}{6} = 0$$

The *expected value* of a discrete random variable X is:

$$E(X) = \sum_{x} xp(x)$$

if the sum exists (actually it has to converge absolutely.)

expected number of coin tosses to first H

Denote by X the number of coin tosses until the first H. The pmf is (new version!):

$$p(x) = \left(1 - \frac{1}{2}\right)^{x-1} \frac{1}{2}$$

for $x \in \{1, 2, 3, \ldots\}$. For a moment replace $\frac{1}{2}$ with p.

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} x(1-p)^{x-1} p \\
&= p \sum_{x=0}^{\infty} x(1-p)^{x-1} \\
&= p \sum_{x=0}^{\infty} \frac{d}{dp} \left(-(1-p)^x\right) \\
&= p \frac{d}{dp} \left(-\sum_{x=0}^{\infty} (1-p)^x\right) = p \frac{d}{dp} \left(-\frac{1}{1-(1-p)}\right) = \frac{1}{p}
\end{aligned}$$

generalization of “tosses to first head”

A factory makes a defective item with probability p (per item) with $0 < p < 1$. What is the expected number of items until the first defective item?

Denote the number of items by X. The pmf of X will be:

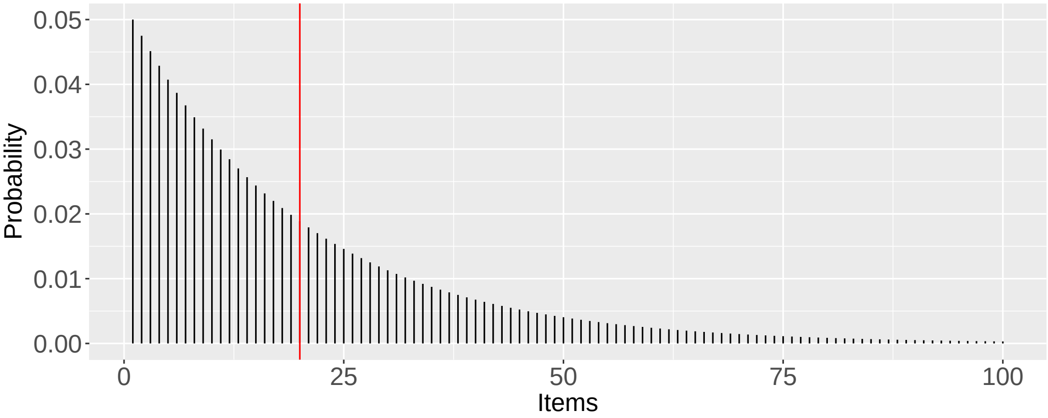
$$p(x) = (1 - p)^{x-1}\, p$$

for $x \in \{1, 2, 3, \ldots\}$.

According to the previous slide, $E(X) = \frac{1}{p}$

graphical view of expected value

Suppose $p = 0.05$. Then $E(X) = 20$. The expected value is the "physical" balance point of the pmf (a.k.a. the *first moment*)



expected value non-example

Let X have the following pmf (!):

$$p(x) = \frac{6}{\pi^2 x^2}, x \in \{1, 2, 3, \ldots\}$$

X does not have an expected value, because $\sum_x xp(x)$ does not converge.

expected value - continuous version

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

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

provided the integral converges (absolutely).

Bus stop example. What is the expected waiting time?

$$\int_{0}^{10} x\frac{1}{10}\,dx = x^2\frac{1}{20}\Big|_{x=0}^{10} = 5$$

another continuous example; plus a non-example

Consider $f(x) = 2x^{-3}$ for $x > 1$, and 0 otherwise. This is a valid density.

Suppose X has density $f(x)$. Then:

$$E(X) = \int_1^\infty x\,2x^{-3}\,dy = \left[-2x^{-1}\right]_{x=1}^\infty = 2$$

But suppose Y has density $f(y) = y^{-2}$ on $y > 1$, and 0 otherwise. Then:

$$\int_1^\infty y\,y^{-2}\,dy = [\log y]_{y=1}^\infty$$

so $E(Y)$ does not exist.

expected values of functions of random variables

Denote by X your outcome after a play of BIG MONEY. Then...

BIG MONEY—*after a few schnapps version*.

We adopt fake German accents and up the game.

Roll	1, 2	3	4, 5	6
Outcome \$	-200	0	100	200
Probability	2/6	1/6	2/6	1/6

Denote by Y your outcome after a play of this modified game. $E(Y)$ is also zero, using the definition:

$$E(Y) = -200\frac{2}{6} + 0\frac{1}{6} + 100\frac{2}{6} + 200\frac{1}{6} = 0$$

expected values of functions of random variables

The definition requires determining the pmf/pdf of the new random variable, which is not always so easy.

Theorem: For a random variable X and a function $g : \mathbb{R} \to \mathbb{R}$:

$$E(g(X)) = \sum_x g(x)p(x) \qquad \text{discrete} \qquad E(g(X)) = \int_{\infty}^{\infty} g(x)f(x)\,dx \qquad \text{continuous}$$

For example, the outcome Y of BIG MONEY (SCHNAPPS VERSION) is related the the outcome X of BIG MONEY by:

$$Y = 100X$$

So the theorem says $E(Y) = E(100X)$ and the calculation is (technically):

$$E(Y) = [100 \cdot (-2)]\frac{2}{6} + [100 \cdot (0)]\frac{1}{6} + [100 \cdot (1)]\frac{2}{6} + [100 \cdot (2)]\frac{1}{6} = 0$$

$E(\cdot)$ rules

The theorem lets us develop some basic rules.

$$\begin{aligned} E(a+bX) &= \sum_x (a+bx)p(x) \\ &= \sum_x ap(x) + \sum_x bxp(x) \\ &= a\sum_x p(x) + b\sum_x xp(x) \\ &= a + bE(X) \end{aligned}$$

(Continuous “version”proof is the same.)

reminder “theorem”

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

“constant” random variables

From last time: $E(a + bX) = a + bE(X)$. The left had side is a bit of an odd expression.

For convenience we can treat constants as random variables. Formally, we can consider a random variable X that is always, say, some real a no matter what. The pmf of X is:

$$P(X = a) = 1$$

and 0 otherwise.

Informally, for convenience, we dispense with the X notation and just treat the constant a as a random variable, allowing for statements like:

$$E(a) = a$$

expected values and joint distributions

Given X and Y with joint density $f(x, y)$, and given a function $g : \mathbb{R}^2 \to \mathbb{R}$, a sophisticated application of the theorem from last time is:

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy$$

(discrete version is the same, with sums)

first key example

Consider $g(x, y) = x + y$.

$$
\begin{aligned}
E(g(X, Y)) = E(X + Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x f(x, y)\, dy \right] dx + \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f(x, y)\, dx \right] dy \\
&= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x, y)\, dy \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f(x, y)\, dx \right] dy \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy \\
&= E(X) + E(Y)
\end{aligned}
$$

second key example

Suppose X and Y are independent. Consider $g(x, y) = xy$.

$$\begin{aligned}
E(g(X, Y)) = E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\, dx\, dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx \int_{-\infty}^{\infty} y f_Y(y)\, dy \\
&= E(X)E(Y)
\end{aligned}$$

notation

It is common to use μ as a shorter stand-in for $E(X)$.

I'm not so convinced this is a good idea.

measuring variation

BIG MONEY versus BIG MONEY SCHNAPPS VERSION

$E(X) = E(Y) = 0$

Roll	1, 2	3	4, 5	6
BIG MONEY Outcome X \$	-2	0	1	2
BIG MONEY SCHNAPPS Outcome Y \$	-200	0	100	200
Probability	2/6	1/6	2/6	1/6

But Y is clearly a riskier game. The distribution is more spread out, by a factor of 100.

The question is—how to measure this difference in variation?

variance

Recall the sample variance, expressed a little differently:

$$s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 \frac{1}{n-1}$$

This is a weighted sum of squared deviations with weights $w_i = 1/(n-1)$

The theoretical analogue is called the *variance*:

$$\mathrm{Var}(X) = E\left((X - E(X))^2\right) = E\left((X - \mu)^2\right)$$

provided the expectations all exist.

It is also possible to view this as an application of the $E(g(X))$ theorem with $g(x) = (x - \mu)^2$.

examples - BIG MONEY

$$\mathrm{Var}(X) = (-2-0)^2\frac{2}{6} + (0-0)^2\frac{1}{6} + (1-0)^2\frac{2}{6} + (2-0)^2\frac{1}{6} = \frac{14}{6}$$

Schnapps version:

$$\mathrm{Var}(Y) = (-200-0)^2\frac{2}{6} + (0-0)^2\frac{1}{6} + (100-0)^2\frac{2}{6} + (200-0)^2\frac{1}{6} = \frac{140000}{6}$$

how to actually calculate variance

Using the two rules $E(a+bX) = a + bE(X)$ and $E(X+Y) = E(X) + E(Y)$ we can derive the better way to calculate variance:

$$\begin{aligned}
\mathrm{Var}(X) &= E\Big((X-\mu)^2\Big) \\
&= E\Big(X^2 - 2X\mu + \mu^2\Big) \\
&= E\Big(X^2\Big) - 2E(X)\mu + \mu^2 \\
&= E\Big(X^2\Big) - \mu^2
\end{aligned}$$

Or as I prefer: $\mathrm{Var}(X) = E(X^2) - E(X)^2$

another variance example

Gas pipes revisited:

	X			
Y	1	1.5	1.75	Marginal
0.5	0.075	0.100	0.150	0.325
1	0.110	0.080	0.090	0.280
2	0.160	0.140	0.095	0.395
Marginal	0.345	0.320	0.335	1.000

$$E(X) = 1 \cdot 0.345 + 1.5 \cdot 0.320 + 1.75 \cdot 0.335 = 1.41125$$

$$E(X^2) = 1^2 \cdot 0.345 + 1.5^2 \cdot 0.320 + 1.75^2 \cdot 0.335 = 2.0909375$$

$$\mathrm{Var}(X) = 0.0993109$$

variance non-example

Reconsider X with density $f(x) = 2x^{-3}$ for $x > 1$, and 0 otherwise. Last time we found $E(X) = 2$.

But X does not have a variance, because $E(X^2)$ does not exist.

notation; standard deviation

It is common to use σ^2 as a shorter stand-in for $\text{Var}(X)$.

I'm not so convinced this is a good idea.

Recall the "unit" problem with the sample variance s^2; hence the use of the sample standard deviation $s = \sqrt{s^2}$.

The theoretical analogue is the *standard deviation*, defined as:

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

It is common to use σ as a shorter stand-in for $\text{SD}(X)$.

I will leave to you to guess as to whether or not I am convinced this is a good idea.

variance rules

$$\text{Var}(a + bX) = b^2\text{Var}(X)$$

What about $\text{Var}(X + Y)$?

variance rules

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

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2E((X - EX)(Y - EY))$$

$E((X - EX)(Y - EY))$ is called the *covariance* of X and Y, or $\mathrm{Cov}(X, Y)$, or σ_{XY}

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)\,E(Y)$$

So, $X \perp Y$ implies Cov(X,Y) $= 0$, in which case the variance of the sum is the sum of the variances.

covariance

$\text{Cov}(X, Y)$ is a numerical summary of a certain kind of relationship that might exist between two distributions.

It is a measure of *linear* relationship, in the following sense(s):

1. If:
   - with high probability, larger values of X and Y happen at the same time, and
   - with high probability, smaller values of X and Y happen at the same time, then:
     - with high probability $(X - EX)(Y - EY)$ will be **positive**.
2. If:
   - with high probability, larger values of X and smaller values of Y happen at the same time, and
   - with high probability, smaller values of X and larger values of Y happen at the same time, then:
     - with high probability $(X - EX)(Y - EY)$ will be **negative**.

correlation coefficient

Covariance is in the unit of the product of the units of X and Y, so its magnitude is uninformative.

A scale-free version is called *correlation coefficient* (or just “correlation”):

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Positive and negative correlation mean the same as positive and negative covariance; in addition, the correlation of different pairs of distributions can be compared. Also:

$$-1 \leqslant \rho \leqslant 1$$

discrete example - positive correlation close to 1

		X		
Y	-1	0	1	Marginal
-1	0.30	0.02	0.01	0.33
0	0.02	0.30	0.02	0.34
1	0.01	0.02	0.30	0.33
Marginal	0.33	0.34	0.33	1.00

$E(X) = E(Y) = 0$, so $\text{Cov}(X, Y) = E(XY)$.

In this case (tedious exercise) $\rho = 0.879$.

The probability is strongly concentrated along the "$X = Y$ diagonal".

discrete example - negative correlation close to 1

	X			
Y	-1	0	1	Marginal
-1	0.01	0.02	0.30	0.33
0	0.02	0.30	0.02	0.34
1	0.30	0.02	0.01	0.33
Marginal	0.33	0.34	0.33	1.00

In this case (tedious exercise) $\rho = -0.879$.

The probability is strongly concentrated along the "$X = -Y$ diagonal".

discrete exmaple - $X \perp Y$

		X		
Y	-1	0	1	Marginal
-1	0.1089	0.1122	0.1089	0.33
0	0.1122	0.1156	0.1122	0.34
1	0.1089	0.1122	0.1089	0.33
Marginal	0.33	0.34	0.33	1.00

X and Y are indepedent (can be tediously verified.)

Easy to show $E(XY) = 0$, so that $\rho = 0$.

discrete example - $\rho = 0$ but very much not independent!

		X		
Y	-1	0	1	Marginal
-1	0.00	0.00	0.25	0.25
0	0.50	0.00	0.00	0.50
1	0.00	0.00	0.25	0.25
Marginal	0.50	0.00	0.50	1.00

X and Y are not independent.

But $E(XY) = 0$, so $\rho = 0$.

standard deviation as an absolute property of a distribution

mean and SD aren't unique, but they do say something

$E(X)$ and $E(X^2)$ 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 $X \geqslant 0$ has expected value $E(X)$, then:

$$P(X \geqslant t) \leqslant \frac{E(X)}{t}.$$

Theorem (Chebyshev): If $\text{Var}(X) = \sigma^2$ and $E(X) = \mu$:

$$P(|X - \mu| \geqslant t) \leqslant \frac{\sigma^2}{t^2}$$

Equivalently:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geqslant 1 - \frac{1}{k^2}$$

example - “uniform on (0,1)”

Suppose X has the density $f(x) = 1$ on $0 < x < 1$ and 0 otherwise.

Then $E(X) = \frac{1}{2}$ and $\text{Var}(X) = \frac{1}{12}$

Various applications of Markov’s and Chebyshev’s inequality for this example show how weak they really are — mainly useful in theory than in practice.

a (strange) new way to completely characterize the distribution of a random variable (note: section 7.3 of book)

what is important about a random variable?

One of the main points of this course so far: 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 $F(x) = P(X \leqslant x)$ (you take on faith that this does the job)
2. (Discrete only) Probability mass functions $p(x)$ (equivalent to cdf)
3. (Continuous only) Probability density functions $f(x)$ (equivalent to cdf)

We will use whichever one is most convenient for a given situation.

mean, variance, and “moments”

$E(X)$ gives a little bit of information about a random variable.

$\text{Var}(X) = E(X^2) - E(X)^2$ gives a little bit *more* information about a random variable.

(c.f. Markov’s and Chebyshev’s inequalities.)

Definition: For integers $k \geqslant 0$, $E\left(X^k\right)$ (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.

the complete moment sequence *characterizes* a distribution

If turns out 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 sequence *usually* gives a *characterization* of the distribution of X.

(I'll tell you how to know *when* this "sometimes" is, momentarily.)

But carrying an around infinite sequence is not convenient—they need to be neatly packaged. Here is the trick:

$$\begin{aligned} M_X(t) &= 1 + E(X)\frac{t}{1!} + E\left(X^2\right)\frac{t^2}{2!} + E\left(X^3\right)\frac{t^3}{3!} + E\left(X^4\right)\frac{t^4}{4!}\cdots \\ &= E\left(1 + Xt + X^2\frac{t^2}{2!} + X^3\frac{t^3}{3!} + X^4\frac{t^4}{4!} + \cdots\right) \\ &= E\left(e^{tX}\right) \end{aligned}$$

moment “generating function”

Definition: $M_X(t)$ is called the *moment generating function*, or mgf, for X (small print: has to converge on an interval containing 0.)

Important fact: the mgf (if it exists) characterizes the distribution of X.

Its less important use is that it can be used to calculate moments:

$$\begin{aligned}
\frac{d^r}{dt^r} M_X(t)\Big|_{t=0} &= \frac{d^r}{dt^r}\left[\int_{-\infty}^{\infty} e^{tx} f(x)\,dx\right]_{t=0} \\
&= \int_{-\infty}^{\infty} \frac{d^r}{dt^r} e^{tx} f(x)\,dx\Big|_{t=0} \\
&= \int_{-\infty}^{\infty} x^r e^{tx} f(x)\,dx\Big|_{t=0} = \int_{-\infty}^{\infty} x^r f(x)\,dx = E(X^r)
\end{aligned}$$

mgf example

Reconsider the generic "observe until the first defective item" example, in which a defective item is produced with probability p. The pmf of the number of items X is:

$$p(x) = (1-p)^{x-1}p \text{ for } x \in \{1, 2, 3, \ldots, \}$$

The mgf of X is:

$$\begin{aligned} M_X(t) = E\left(e^{tX}\right) &= \sum_{x=1}^{\infty} e^{tx}(1-p)^{x-1}p \\ &= pe^t \sum_{x=1}^{\infty} [e^t(1-p)]^{x-1} \\ &= \frac{pe^t}{1-(1-p)e^t} \end{aligned}$$

Tedious calculus shows $M'_X(t) = \frac{pe^t}{(1-(1-p)e^t)^2}$, which when evaluated at 0 is $\frac{1}{p}$. (I hope!)

the more important use of of mgfs

Recall that if X and Y are independent, $E(XY) = E(X)E(Y)$. This can extend easily to:

$$E(g(X)h(Y)) = E(g(X))E(h(Y))$$

Now, consider the random variable $X + Y$ with X and Y independent. What could be said about the *distribution* of $X + Y$? This is a difficult problem!

But the following result will be useful over the next few weeks:

$$M_{X+Y}(t) = E\left(e^{t(X+Y)}\right) = E\left(e^{tX}e^{tY}\right) = E\left(e^{tX}\right)E\left(e^{tY}\right) = M_X(t)M_Y(t)$$

distributions with names, because they are used all the time to model actual things

a few conventions

Some families of random variables are common enough that they get their own names.

“The Foo Distribution”

Typically the Foo Distribution will only be specified up to some constants that are called parameters. The parameters often have some meaning or another.

Notation: $X \sim \text{Foo}(\delta, \nu)$

This is pronounced: “X has a Foo distribution with parameters δ and ν.”

Finally, a collection of random variables is often called a “process”. We will examine two processes that are central to probability modeling.

the Bernoulli distributions

We've seen this one a few times. X has two outcomes: 0 and 1. We say $X \sim$ Bernoulli(p). The parameter is $p = P(X = 1)$.

pmf: $p(x) = p^x(1-p)^{1-x}, \quad x \in 0, 1$

Sometimes $1 - p = q$ is used when convenient.

$E(X) = (0)(1-p) + (1)(p) = p$

For variance, use $E(X^2) = 0^2(1-p) + 1^2 p = p$, so that
$\text{Var}(X) = p - p^2 = p(1-p) = pq$

$SD(X) = \sqrt{p(1-p)}$

$M_X(t) = E(e^{tX}) = e^{t \cdot 0}q + e^{t \cdot 1}p = q + e^t p$

Bernoulli process

The Bernoulli(p) distributions are used as models for anything with two possible outcomes.

There is usually more than one at a time. A *Bernoulli process* is a sequence of independent Bernoulli(p) random variables with the same p:

$$X_1, X_2, X_3, X_4 \ldots$$

Often each X_i is called a (Bernoulli) *trial*, or an *experiment*.

after n trials, how many 1's?

Fix the number of Bernoulli trials at n, and let X be the count of the 1's that occurred. What is the *distribution* of X?

First, X could be any integer between 0 and n.

To illustrate a probability calculation, fix $n = 4$; we'll consider $P(X = 2)$.

Trial outcomes	Probability
1 1 0 0	$p^2(1-p)^2$
1 0 1 0	$p(1-p)p(1-p) = p^2(1-p)^2$
1 0 0 1	$p^2(1-p)^2$
0 1 1 0	$p^2(1-p)^2$
0 1 0 1	$p^2(1-p)^2$
0 0 1 1	$p^2(1-p)^2$

the Binomial(n, p) distributions

There are 6 ways to get 2 1s in 4 trials, so $P(X = 2) = 6p^2(1-p)^2$.

In general, the number of ways to get k 1s in n trials is:

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

If X is the number of 1s in n Bernoulli(p) trials, we say:

$$X \sim \text{Binomial}(n, p)$$

pmf: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k \in \{0, 1, \ldots, n\}$

Is this actually a valid pmf? Yes, if you recall the *Binomial theorem*:

$$(a+b)^n = \sum_{k=0} \binom{n}{k} a^k b^{n-k}$$

Let $a = p$ and $b = 1 - p$.

the Binomial(n, p) distributions

Model for the number of 1s after n trials of a Bernoulli(p) process.

pmf: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k \in \{0, 1, \ldots, n\}$

Path to $E(X)$ is easiest through the mgf:

$$M_X(t) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k q^{n-k} = (q + pe^t)^n$$

Theorem: Let X be the sum of n independent Bernoulli(p) random variables $X_1, \ldots, X_n$. Then $X \sim$ Binomial(n, p).

Proof:

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t) = (q + pe^t)^n$$

the Binomial(n, p) distributions

So:

$$E(X) = E(X_1) + \cdots + E(X_n) = np$$

and

$$\text{Var}(X) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = np(1-p)$$

$$SD(X) = \sqrt{np(1-p)}$$

number of Bernoulli trials until the first 1—Geometric(p)

We've done most of the work on this distribution.

$$p(y) = (1-p)^{y-1}p, \quad y \in \{1, 2, 3, \ldots\}$$

$$E(Y) = \frac{1}{p}$$

$$M_Y(t) = \frac{pe^t}{1 - qe^t}$$

Variance is inevitably tedious. $E(Y^2) = \frac{d^2}{dt^2} M_Y(t)\Big|_{t=0} = \frac{2-p}{p^2}$, resulting in $\text{Var}(Y) = \frac{q}{p^2}$

We say $Y \sim \text{Geometric}(p)$, due to the geometric rate of decay in the pmf.

cdf and the “reliability function” for a Geometric(p)

The cdf for $Y \sim \text{Geometric}(p)$ comes in handy sometimes.

$$F_Y(y) = \begin{cases} 0 & : y < 1 \\ 1 - (1-p)^k & : k \leqslant y < k+1 \end{cases}$$

The *reliability* or *survival* function of a random variable is defined as

$$R(x) = P(X > x) = 1 - P(X \leqslant x) = 1 - F(x).$$

(In class I said “left continuous”; in fact it is right continuous.)

For Y in this case (corrected from class):

$$R_Y(y) = \begin{cases} 1 & : y < 1 \\ (1-p)^k & : k \leqslant y < k+1 \end{cases}$$

the memorylessess of a Geometric(p) distribution

Bernoulli process is the discrete model for what I will (vaguely) refer to as "complete randomness", which will come to embody these (vague) ideas:

- independence (of trials)
- homogeneity (p doesn't change)
- memoryless (information does not increase with number of trials)

The memorylessness of the Bernoulli process can be seen through a property of the Geometric(p) distribution. Suppose $Y \sim$ Geometric(p).

$$\begin{aligned} P(Y > j + k | Y > j) &= \frac{P(Y > j + k, Y > j)}{P(Y > j)} \\ &= \frac{P(Y > j + k)}{P(Y > j)} = \frac{(1-p)^{j+k}}{(1-p)^j} = (1-p)^k = P(Y > k) \end{aligned}$$

number of Bernoulli trials until the r^{th} 1

Denote by W the number of trials in a Bernoulli(p) process until the r^{th} 1 occurs.

We'll say W has a "Negative Binomial" distribution, or $W \sim \text{NegBin}(r, p)$.

e.g. consider $r = 4$ and $W = 11$. $P(W = 11)$ can be derived as follows:

Trial outcomes	Probability
1 1 0 0 0 0 0 1 0 0 1	$(1-p)^7 p^3 p$
1 0 1 0 0 1 0 0 0 0 1	$(1-p)^7 p^3 p$
$\vdots$	$\vdots$

It will always be 3 1s out of 10 trials, follow by a 1. So the probability is:

$$P(W = 11) = \binom{11-1}{4-1}(1-p)^{11-4}p^4$$

In general: $P(W = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r, \quad k \in \{r, r+1, r+2, \ldots\}$

NegBin(r, p)

The name comes from this version of the Binomial theorem with negative exponent:

$$(1-a)^{-r} = \sum_{k=r}^{\infty} \binom{k-1}{r-1} a^{k-r}$$

which can be used to verify that the pmf is legit.

Next up—mgf.

$$\begin{aligned} M_W(t) &= \sum_{k=r}^{\infty} e^{tk} \binom{k-1}{r-1} (1-p)^{k-r} p^r \\ &= (pe^t)^r \sum_{k=r}^{\infty} \binom{k-1}{r-1} (qe^t)^{k-r} \\ &= \left(\frac{pe^t}{1-qe^t} \right)^r \end{aligned}$$

NegBin(r, p)

Theorem: Let W be the sum of r independent random variables $Y_1, \ldots, Y_r$ with Geometric(p) distributions. Then $W \sim$ NegBin(r, p).

Proof: $M_{Y_1+\cdots+Y_r}(t) = M_{Y_1}(t) \cdots M_{Y_r}(t) = \left(\frac{pe^t}{1-qe^t}\right)^r$

So:

$$E(W) = E(Y_1) + \cdots + E(Y_r) = \frac{r}{p}$$

and

$$\text{Var}(W) = \text{Var}(Y_1) + \cdots + \text{Var}(Y_r) = \frac{rq}{p^2}$$

the hypergeometric distributions

Reminder: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of ways to choose k items out of n, without replcement.

Quality control methods often deal with this sort of situation:

100 items are produced in which 5 are defective. A sample of 10 is selected. Denote by X the number of defective items out of the 10 selected.

X could take on integer values between 0 and 5 inclusive, with probabilities:

$$P(X = x) = \frac{\binom{5}{x}\binom{95}{10-x}}{\binom{100}{10}} = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

N is the population size. n is the sample size. k is the number of "defective" (in this example.)

and we say X has a Hypergeometric distribution with parameters N, n, and k.

the hypergeometric distributions

I'm not sure where the name comes from. There is no natural relationship to the Bernoulli process, but there is a tangential one...

Hypergeometric is a "sampling without replacement" version of the Binomial. If N and n are both large, it turns out:

$$\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \approx \binom{n}{x}\left(\frac{k}{N}\right)^{x}\left(1-\frac{k}{N}\right)^{n-x}$$

All calculations involving these distributions are tedious. It can be shown (no easy way):

$$E(X) = n\frac{k}{N} \qquad \operatorname{Var}(X) = n\frac{k}{N}\left(1-\frac{k}{N}\right)\left(\frac{N-n}{N-1}\right)$$

“binomial process”

Let’s put a time scale on a Bernoulli process. One trial will now happen in every unit of time Δ (called a “frame”), and successes happen now with a “rate” λ per unit of time.

The result of this re-jigging is a sequence $X_1, X_2, X_3, \ldots$ of random variables that are independent with Bernoulli($\lambda\Delta$) distribution.

Consider the cumulative sums $X(t)$ of this Bernoulli process:

$$\begin{aligned} X(1) &= X_1 \\ X(2) &= X_1 + X_2 \\ &\vdots \\ X(n) &= \sum_{i=1}^{n} X_i \\ &\vdots \end{aligned}$$

“binomial process”

$X(t)$ is non-negative, integer valued, non-decreasing sequence of random variables. Such a sequence is called a *counting process*.

This particular counting process is called a *binomial process*. It counts the number of 1s in the first n frames—equivalent to the number of 1s from time $t = 0$ to time $t = n\Delta$.

(Assume the Bernoulli trial occurs at the start of the frame.)

$X(t)$ has these properties:

- $X(0) = 0$
- $X(t) \sim \text{Binomial}\left(n = \frac{t}{\Delta}, p = \lambda\Delta\right)$.
- $X(t)$ increments by 0 or 1 per frame.
- Number of frames between increments is Geometric(p).
- Given times $s_1 < t_1 < s_2 < t_2$, the differences $X(t_1) - X(s_1)$ and $X(t_2) - X(s_2)$ are *independent* random variables.
- $E(X(t)) = np = \frac{t}{\Delta}\lambda\Delta = \lambda t$

pass to the limit

Fix a time t. Keep $E(X(t)) = np = \lambda t$ constant, so that $p = \lambda t / n$.

The goal is to find the limit of $P(X(t) = k)$ as $n \to \infty$ for any $k \geqslant 0$.

For any fixed n, $X(t) \sim \text{Binomial}\left(n, \frac{\lambda t}{n}\right)$

pass to the limit

$$\begin{aligned}
\lim_{n\to\infty} P(X(t)=k) &= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{n-k} \\
&= \frac{(\lambda t)^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1-\frac{\lambda t}{n}\right)^{n} \left(1-\frac{\lambda t}{n}\right)^{-k} \\
&= \frac{(\lambda t)^k}{k!} \lim_{n\to\infty} \underbrace{\frac{n\cdot(n-1)\cdot(n-2)\cdots(n-k+1)}{n\cdot n\cdot n\cdots n}}_{k\text{terms}} \left(1-\frac{\lambda t}{n}\right)^{n} \left(1-\frac{\lambda t}{n}\right)^{-k} \\
&= \frac{(\lambda t)^k}{k!} e^{-\lambda t}
\end{aligned}$$

Poisson process/distribution

Nothing to do with fish. Named after some French guy.

A Poisson process $N(t)$ is a counting process which counts the number of "events" that happen inside the time interval $[0, t]$, subject to the following:

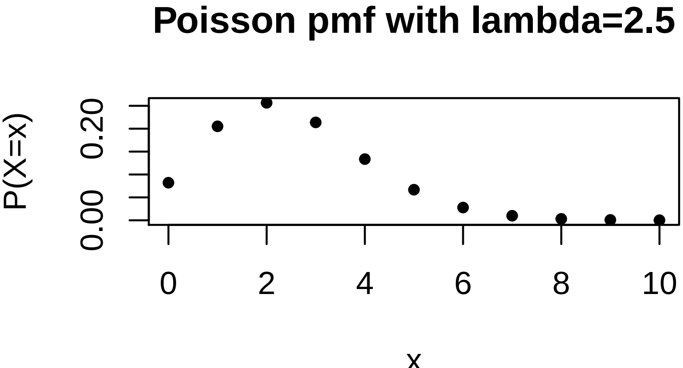
- $N(0) = 0$
- Given times $s_1 < t_1 < s_2 < t_2$, the differences $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are *independent* random variables.
- $P(N(t) - N(s) = k) = \frac{(\lambda(t-s))^k}{k!} e^{-\lambda(t-s)}$ for $0 \leqslant s < t$

The Poisson process is a common model for events that happen "completely randomly" in time (to be further discussed)

Poisson distribution

There is also the closely related Poisson distribution, in which the t is not explicitly used. We say $X \sim \text{Poisson}(\lambda)$ if it has pmf:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for } k \in \{0, 1, 2, \ldots\}$$



Poisson examples

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 or why not might a Poisson process model be suitable here?

other Poisson distribution properties

Suppose $X \sim \text{Poisson}(\lambda)$. Then (using s as the meaningless "dummy" variable):

$$M_X(s) = E(e^{sX}) = \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s \lambda)^k}{k!} = e^{-\lambda} e^{e^s \lambda} = e^{\lambda(e^s - 1)}$$

Then it's easy to show $E(X) = \lambda$ and $\text{Var}(X) = \lambda$.

Using full-blown Poisson process notation, we have that the expect value and the variance of the number of events in $[s, t]$ are both $\lambda(t - s)$.

Poisson - when did the event happen?

In a Poisson process we might have observed that $N(t) = 1$. (We will condition on this fact throughout.)

When inside the interval $[0, t]$ did that single event occur?

The time when the event happened is a continuous random variable, which we can call X. What is it's distribution?

We can derive its cumulative disribution function from first principles. We know $P(X \leqslant x | N(t) = 1) = 0$ for $x \leqslant 0$ and $P(X \leqslant x | N(t) = 1) = 1$ for $x > t$.

cdf of "the time when that one thing happened" - cont'

Between 0 and t we have:

$$\begin{aligned}
P(X \leqslant x | N(t) = 1) &= 1 - P(X > x | N(t) = 1) \\
&= 1 - P(N(x) = 0 | N(t) = 1) \\
&= 1 - \frac{P(N(x) = 0, N(t) = 1)}{P(N(t) = 1)} \\
&= 1 - \frac{P(N(x) = 0, N(t) - N(x) = 1)}{P(N(t) = 1)} \\
&= 1 - \frac{P(N(x) = 0)P(N(t) - N(x) = 1)}{P(N(t) = 1)} \\
&= 1 - \frac{e^{-\lambda x}\lambda(t - x)e^{-\lambda(t-x)}}{\lambda t e^{-\lambda t}} \\
&= 1 - \frac{t - x}{t} = \frac{x}{t}
\end{aligned}$$

density of "the time when that one thing happened"

Putting it all together, the cdf of the time X when the event occurred given $N(t) = 1$ is:

$$F_X(x) = \begin{cases} 0 & : x \leqslant 0 \\ \frac{x}{t} & : 0 < x \leqslant t \\ 1 & : x > t \end{cases}$$

So the density is:

$$f_X(x) = \frac{d}{dx} F_X(x) = \begin{cases} \frac{1}{t} & : 0 < x \leqslant t \\ 0 & : \text{ otherwise} \end{cases}$$

The density is flat between 0 and t. We call this a "uniform distribution" between 0 and t.

the uniform distributions

The random variable X, the result of “picking a real number at random between a and b”, is modeled using a flat density:

$$f(x) = \begin{cases} \frac{1}{b-a} & : a < x < b \\ 0 & : \text{ otherwise} \end{cases}$$

Easy to show $E(X) = (a+b)/2$ and $\mathrm{Var}(X) = (b-a)^2/12$.

mgf is easy to determine, but not really useful for anything.

We say $X \sim \mathrm{U}[a, b]$, with $\mathrm{U}[0, 1]$ an important special case.

Poisson approximation to binomial

The limit $P(X(t) = k) \to P(N(t) = k)$ converges very fast, which means difficult Binomial calculations can be approximated very accurately using a Poisson probability calculations.

This was great in the 1960s, but not so important now. There is a discussion in the textbook and a handful of textbook exercises we'll call "optional" if you are intested.

waiting time to the first event of a Poisson process

In a Poisson process with rate λ, the first event will happen at some random time X. What is the distribution of X?

We can derive the cdf from first principles by observing the following:

$$P(X > t) = P(N(t) = 0)$$

The cdf is 0 for $t < 0$. Otherwise:

$$F_X(t) = P(X \leqslant t) = 1 - P(X > t) = 1 - P(N(t) = 0) = 1 - \frac{(\lambda x)^0 e^{-\lambda t}}{0!} = 1 - e^{-\lambda t}$$

The density is therefore 0 when $t < 0$ and otherwise:

$$f_X(t) = \lambda e^{-\lambda t}$$

the exponential distributions

We say X has an exponential distribution with rate $\lambda > 0$ when it has density $f_X(t) = \lambda e^{-\lambda t}$ for $t > 0$.

The m.g.f. is:

$$M_X(s) = \int_0^\infty e^{st} \lambda e^{-\lambda t}\, dt = \frac{\lambda}{\lambda - s}$$

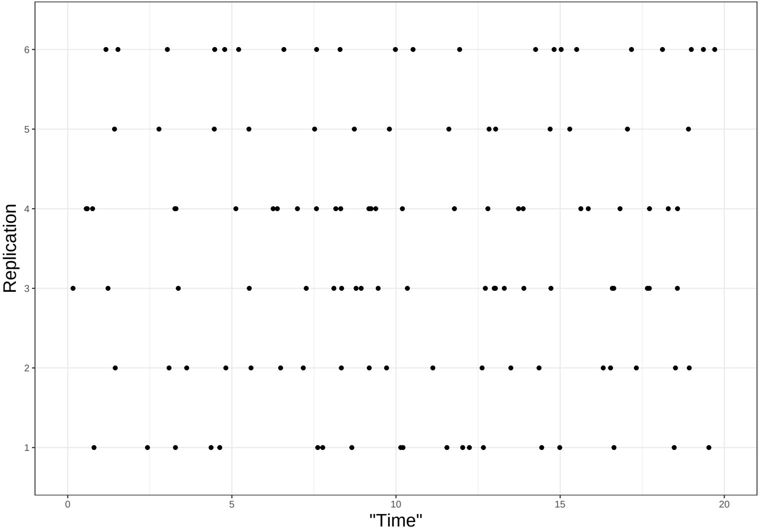
The mean and variance are $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$

Exponential distributions have the *memoryless* property, which is a crucial aspect of Poisson process as a model for "complete randomness".

The Exponential distributions are the *only* (continuous) memoryless distributions.

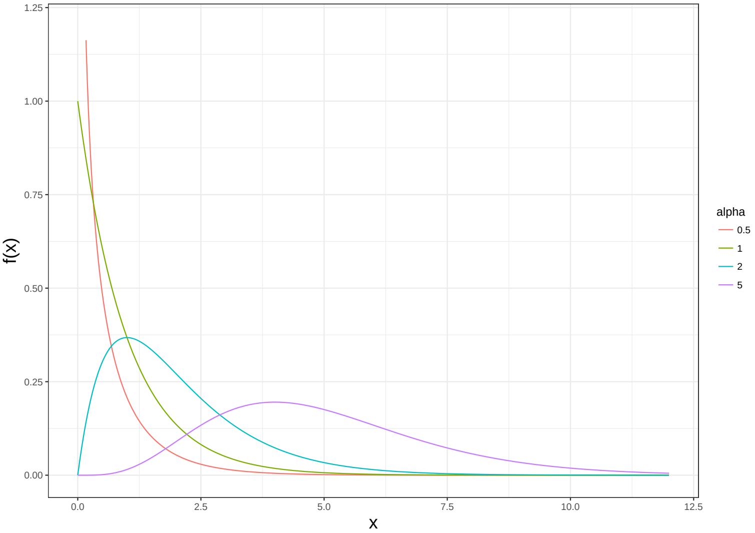
DISASTROUS TREND SHOCKER PANIC HEADLINE

complete randomness is hard for humans



Answer: 1, 3, 4

pictures of some Gamma(α, 1) densities



properties of gamma distributions

If $X \sim \text{Gamma}(\alpha, \lambda)$, its moment generating function can be found to be:

$$M_X(s) = \left(\frac{\lambda}{\lambda - s}\right)^{\alpha}$$

(so the mean and variance are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $\text{Exp}(\lambda)$. What is the distribution of $X = X_1 + X_2 + \cdots + X_n$?

Using the m.g.f. argument it is clear that $X \sim \text{Gamma}(n, \lambda)$, which makes sense in the Poisson process context.

waiting time to the 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 $F(t) = 1 - P(X > t)$, and $\{X > t\}$ *is exactly equivalent to* $\{N(t) \leqslant n - 1\}$, so we can derive the cdf for X.

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

So the density is (a long telescoping sum of work later. . .):

$$f(t) = F'(t) = \begin{cases} \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} & : t > 0 \\ 0 & : \text{otherwise.} \end{cases}$$

the gamma distributions

The density is a special class of a larger family of 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 $n \geqslant 1$.

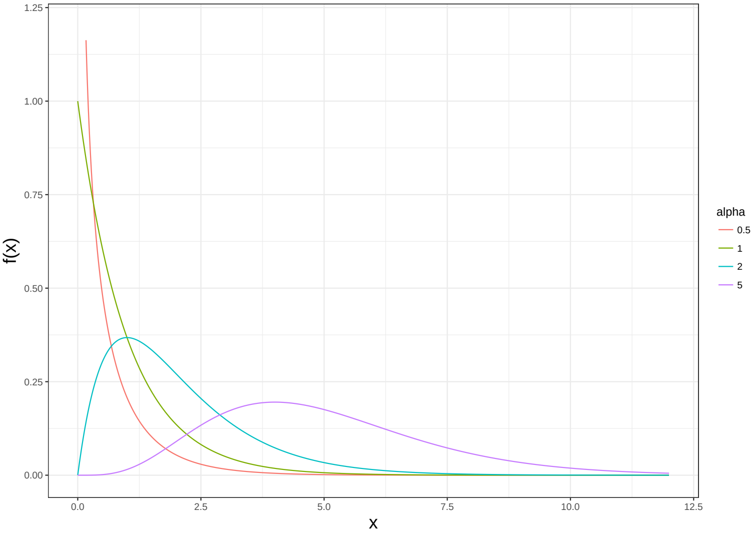
The following function is a valid density for $\alpha > 0$ and $\lambda > 0$:

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

The parameters α and λ are called the *shape* and rate parameters. We say $X \sim \text{Gamma}(\alpha, \lambda)$

$\alpha = 1$ is the special case of $\text{Exp}(\lambda)$.

pictures of some Gamma(α, 1) densities



properties of gamma distributions

If $X \sim \text{Gamma}(\alpha, \lambda)$, its moment generating function can be found to be:

$$M_X(s) = \left(\frac{\lambda}{\lambda - s}\right)^{\alpha}$$

(so the mean and variance are $\frac{\alpha}{\lambda}$ and $\frac{\alpha}{\lambda^2}$)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $\text{Exp}(\lambda)$. What is the distribution of $X = X_1 + X_2 + \cdots + X_n$?

Using the m.g.f. argument it is clear that $X \sim \text{Gamma}(n, \lambda)$, which makes sense in the Poisson process context.

summary of the Bernoulli-Poisson-o-sphere

Starting with a Bernoulli(p) process, we have the following:

What?	Discrete Version	Comments	Continuous Version
Count	Binomial(n, p)	Sum of n Bernoulli(p). Fix $E(X(t)) = np = \lambda t$ fixed $n \to \infty \dots$	$\dots$ Poisson(λt)
Inter-arrival	Geometric(p)	"Memoryless"	Exponential(λ)
Wait for r^{th} event	NegBin(r, p)		Gamma(r, λ)
Look back after 1	"Discrete Uniform"	($\leftarrow$ not done)	Uniform$(0, t)$

a re-parametrization

There was nothing sacred about using λ in the definitions of the exponential and gamma distributions.

Any 1-1 function of a parameter will do. For example, it is common to “parametrize” the exponential and gamma distributions by $\beta = 1/\lambda$ instead.

β is called the “mean” or “scale” parameter, in this case.

This is the book’s parametrization, despite being an engineering book.

But it doesn’t matter.

Although it makes me scarlet with rage.

the “normal” distributions

distributions of sums

We've already seen many examples of things that can be considered as *sums of random variables*.

Binomial

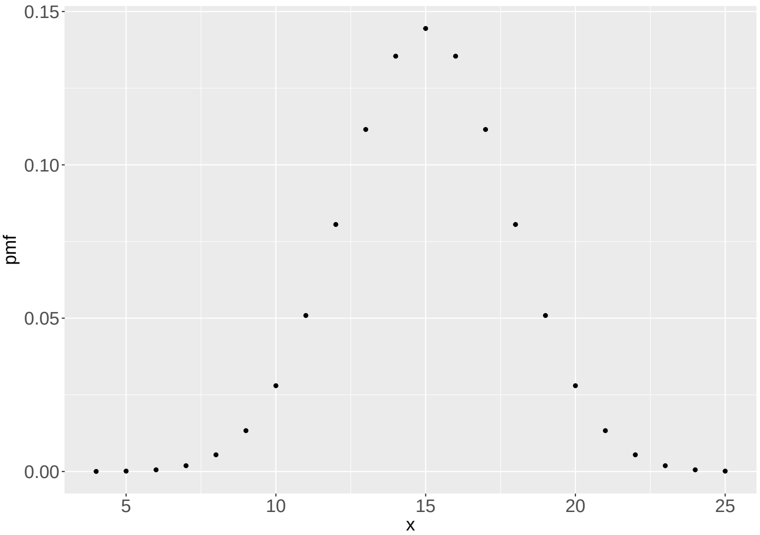
Negative Binomial

Gamma

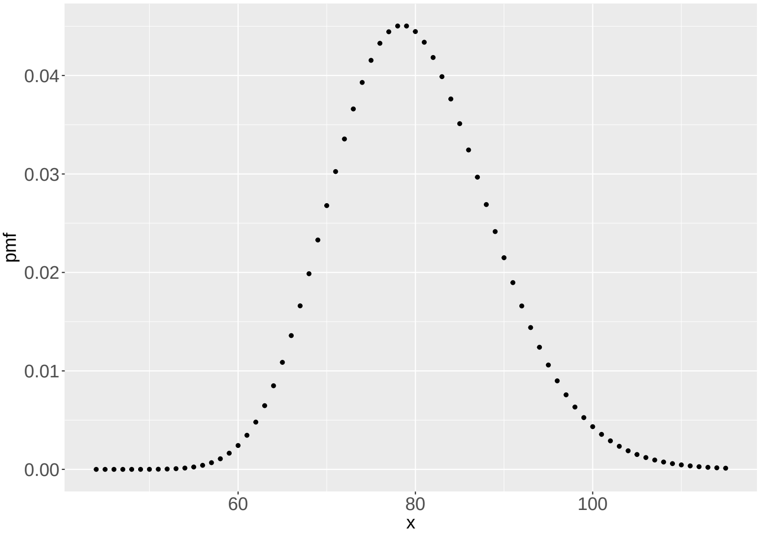
Even Poisson, in the broader sense of "summing up events over an interval"

Let's see what happens with the sum is of not a small number of terms...

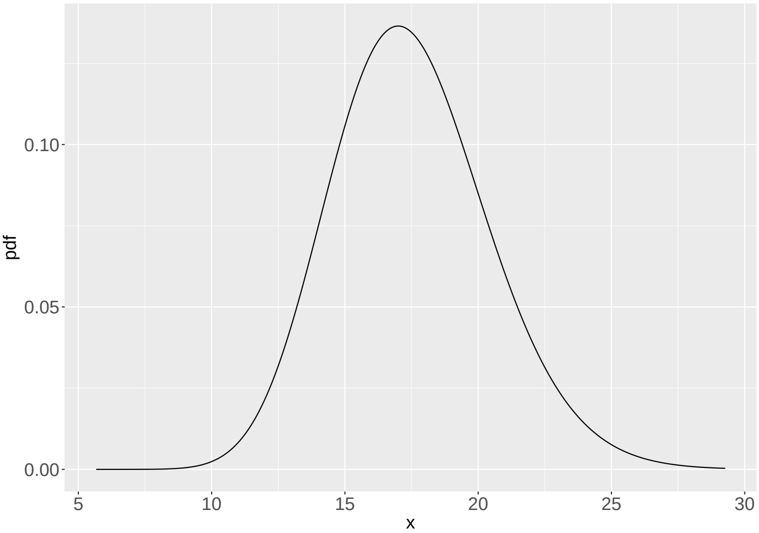
Binomial(30, 0.5)



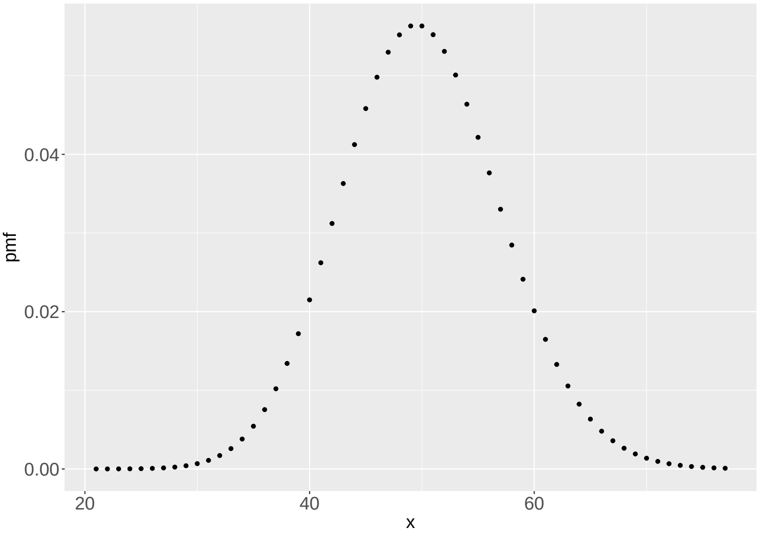
NegBin(40, 0.5)



Gamma(35, 2)



Poisson(50)



normal distributions “in the wild”

Some things actually have normal distributions (symmetric, bell-shaped, no extreme values) in and of themselves.

Such as when the random “thing” is the combination of:

- not a small number of. . .
- . . . roughly equally weighted
- . . . mostly independent. . .
- . . . other random things.

Height, test score, lab measurement, etc.

But this fact undersells the critical importance of the normal distributions.

the normal distributions

We say Z has a "standard" normal distribution, or $Z \sim N(0,1)$, if its density is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

Is this a density?

$E(Z) = 0$ (easy integral) and $\text{Var}(Z) = 1$ (easy integration by parts).

Since $\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz = 1$, the change of variables $z = \frac{x-\mu}{\sigma}$ for any μ and any $\sigma > 0$ shows the following is also a valid density:

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

We say X has a normal distribution with parameters μ and σ, or $X \sim N(\mu, \sigma)$, when it has this density.

the normal distributions

We say Z has a "standard" normal distribution, or $Z \sim N(0,1)$, if its density is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \qquad -\infty < z < \infty$$

Is this a density?

$E(Z) = 0$ (easy integral) and $\text{Var}(Z) = 1$ (easy integration by parts).

Since $\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz = 1$, the change of variables $z = \frac{x-\mu}{\sigma}$ for any μ and any $\sigma > 0$ shows the following is also a valid density:

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

We say X has a normal distribution with parameters μ and σ, or $X \sim N(\mu, \sigma)$, when it has this density.

normal mgf and some implications

Tedious algebra and calculus show (not interesting - see textbook or internet):

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

Suppose $Z \sim N(0,1)$. What is the distribution of $Y = \mu + \sigma Z$ for any μ and $\sigma > 0$?

Answer: $Y \sim N(\mu, \sigma)$.

Suppose $X \sim N(\mu, \sigma)$. What is the distribution of $Z = \frac{X-\mu}{\sigma}$?

Answer: $Z \sim N(0,1)$.

A linear transformation of a normal is normal.

If X_1 and X_2 are normal, what is $X_1 + X_2$?

normal probability calculations

Since the normal density has no anti-derivative (the usual case with functions, BTW), probability calculations are a problem for the computer, or a table of probabilities, which you will need to practice if necessary.

Why only one table when there are many normal distributions?

I'll do a few examples from the `normal_bootcamp.pdf` drill document contained with these notes.

some normal probabilities to remember

The classic plus-or-minus $k\sigma$ probabilities (with exact values):

$$P(\mu - \sigma < X < \mu + \sigma) \approx 0.68 \qquad (0.6826895)$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95 \qquad (0.9544997)$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997 \qquad (0.9973002)$$

Bonus probabilities:

$$P(\mu - 6\sigma < X < \mu + 6\sigma) = 0.999999998027$$

$$P(\mu - 8.5\sigma < X < \mu + 4.5\sigma) = 0.999996602327$$

other distributions

There are lots of other distributions (continuous and discrete) that have many applications, but we'll stop here.

There will be a few more special-purpose distributions specific for data analysis that I'll introduce when necessary.

statistics

the general problem, and its solution

For every probability problem done so far, the distribution was completely known, along with all paramater values.

The problem is: often we encounter one or more distributions that are only partially specified, such as:

- $N(\mu, 2)$, μ unknown.
- $N(\mu, \sigma)$, both parameters unknown.
- "Some distribution that has unknown mean and variance."

etc.

The solution to this problem is to obtain a dataset, and use it to *infer* statements about the underlying distribution.

a probabilistic model for the phrase “to obtain a dataset”

One model for this prospective dataset can be to consider it as a mix of columns of length n where some (or all) of the columns are random.

A random column is headed by random variable (with “the underlying distribution”), and the contents are random variable “copies” of that underlying distribution.

There could be non-random columns with categorical or numerical information.

"Subject ID"	X	Y	"Group ID"	"InputVar"
ID345	X_1	Y_1	A	w_1
ID952	X_2	Y_2	A	w_2
ID826	X_3	Y_3	B	w_3
ID118	X_4	Y_4	B	w_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ID503	X_n	Y_n	A	w_n

the dataset, once observed, is fixed

The model for a dataset is a model for the plan you make to collect it.

Every method of analysis we will discuss is based on this plan to collect.

. . . because once the dataset is collected, there is nothing random about it. It is a fixed rectangle of numbers/etc.

definition of sample | definition of statistic

The basic model for what we'll call a *sample* is a sequence of random variables that are independent with the same distribution (abbreviation: i.i.d.):

$$X_1, X_2, \ldots, X_n$$

This is a column of the dataset.

Some practical generalizations we won't touch: the elements are not independent; the elements do not have the same distribution (e.g. change as a function of time.)

A *statistic* is defined as *a function of the sample.* So, a statistic is a random variable.

example of statistic

The *sample mean* (or sample average):

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

Source of confusion: the function $\overline{x} = \sum_{i=1}^{n} x_i/n$ of a column $\{x_1, \ldots, x_n\}$ in an *observed* dataset is also called “sample mean”, but this is not a random quantity.

I might call $\overline{x}$ the “observed” sample mean, but you really just need to get used to the terminology.

You might have a plan to guess the unknown “true” mean of a random variable X using the sample mean $\overline{X}$ of a sample $X_1, \ldots, X_n$ (whose properties can be studied using probability), and when you actually observe the sample $x_1, \ldots, x_n$ your actual guess will be $\overline{x}$.

more “statisics”

The *sample variance*:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

(source of confusion: the (observed) sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$)

The *sample standard deviation*: $S = \sqrt{S^2}$

More:

- the sum $\sum_{i=1}^{n} X_i$
- the *minimum* $X_{(1)}$
- the *maximum* $X_{(n)}$
- the *sample median* $\tilde{X}$
- the *sample range* $X_{(n)} - X_{(1)}$

“sampling distributions”

Statistics are random variables, so we mainly care about their distributions, which are (unnecessarily) called the “sampling distributions”.

Example, if $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p), the statistic $\sum_i X_i$ has a Binomial(n, p) distribution. (What about $\overline{X}$ in this case?)

Example: if $X_1, \ldots, X_n$ are i.i.d. Exponential(λ), the sample mean $\overline{X}$ has a Gamma($n, n\lambda$) distribution.

Important example: if $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma)\ldots$

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

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

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \dots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum\limits_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum\limits_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

Using the rules for normal distributions we also get:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad\qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

The seemingly impossible task is to determine the distribution of $\overline{X}$ when the underlying distribution is not normal (i.e., almost always.)

basic result for $\overline{X}$ in general

For a sample $X_1, \ldots, X_n$ i.i.d. from *any* distribution with mean μ and variance σ^2, the following are always true:

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

$$\text{Var}\left(\overline{X}\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

full distribution of $\overline{X}$ when sample is normal

For a sample $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma)$ we have:

$$M_{\sum\limits_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(e^{\mu t + \sigma^2 t^2/2}\right)^n = e^{n\mu t + n\sigma^2 t^2/2}$$

so that $\sum\limits_{i=1}^{n} X_i \sim N\left(n\mu, \sqrt{n}\sigma\right)$

Using the rules for normal distributions we also get:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad\qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

The seemingly impossible task is to determine the distribution of $\overline{X}$ when the underlying distribution is not normal (i.e., almost always.)

the actual central limit theorem

For any random variable X with mean μ and variance σ^2, consider a sample $X_1, \ldots, X_n$ from the same distribution.

Now consider the sample average of this sample: $\overline{X}_n$ (because n will be changing below...).

The actual Central Limit Theorem (CLT) says:

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant u\right) = P(Z \leqslant u)$$

$$\lim_{n\to\infty} F_n(u) = F_Z(u)$$

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

the value of the CLT is in the speed of convergence

Useful limit theorems are ones where the convergence is fast—the CLT is such an example.

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant u\right) \approx P(Z \leqslant u)$$

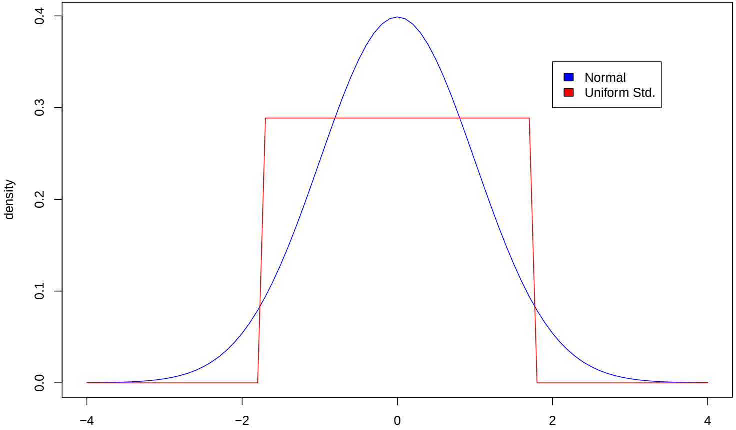
for n “large enough”.

How large? Depends on the shape of the underlying distribution X.

- $n = 2$ would need X normal.
- $n = 10$ good enough for symmetric distributions without outliers.
- $n = 30$ for mildly skewed X.
- $n > 60$ for more skewed.

how large is large enough case I - symmetric/no outliers

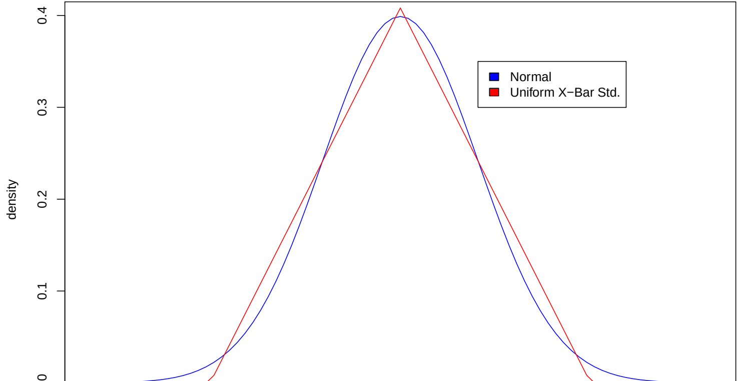
Consider a Uniform[0,1] distribution. The mean is 0.5 and the standard deviation is $1/\sqrt{12}$. Here is a plot of a *standardized* uniform density versus a $N(0,1)$ density:



how large is large enough case I - symmetric/no outliers

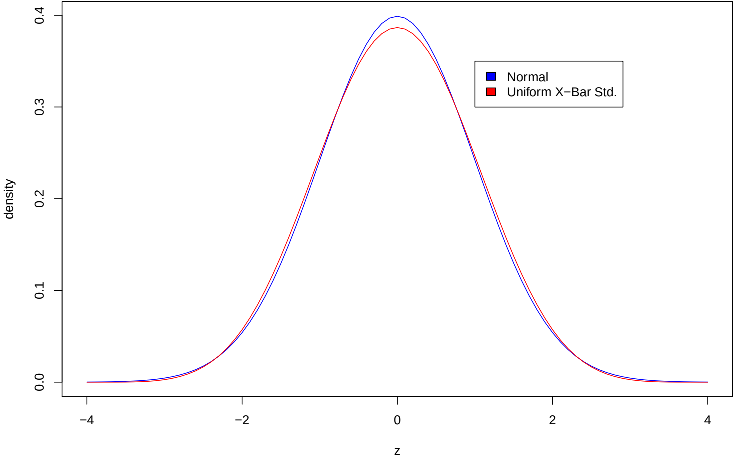
Now consider X_1, X_2 i.i.d. Uniform[0,1], and its sample average $\overline{X}$, which will have mean 0.5 and standard deviation $1/\sqrt{12 \cdot 2}$.

Here is a picture of the density for $\frac{\overline{X}-0.5}{(1/\sqrt{12 \cdot 2})/\sqrt{2}}$, along with the density for $Z \sim N(0,1)$.



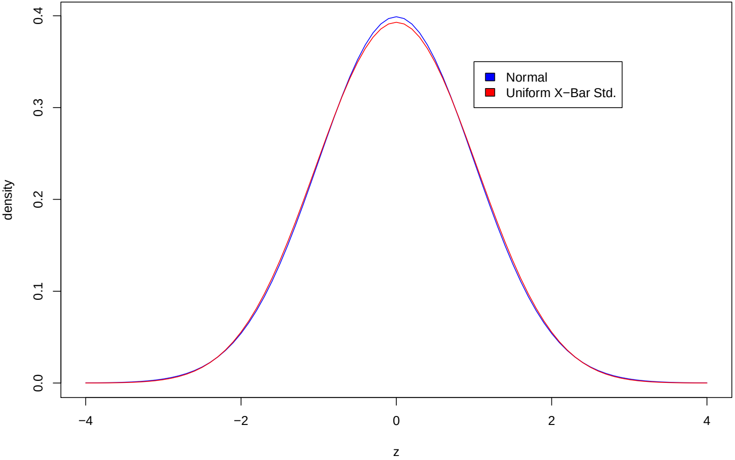
how large is large enough case I - symmetric/no outliers

Same as before but now with X_1, X_2, X_3, X_4, X_5 (i.e. $n = 5$):



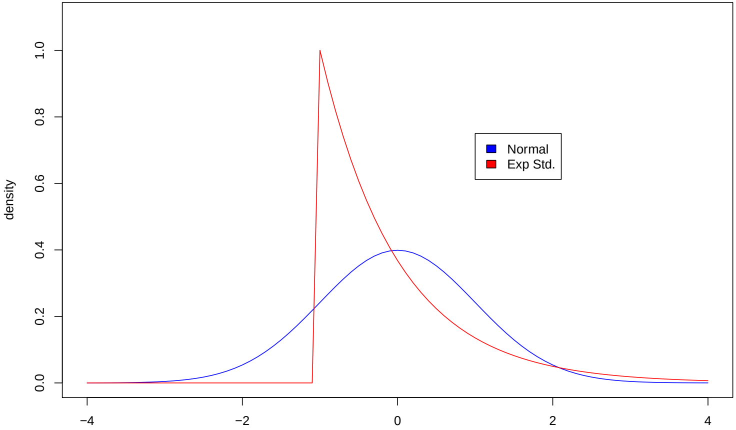
how large is large enough case I - symmetric/no outliers

Now with $n = 10$.



how large is large enough case II - skewed

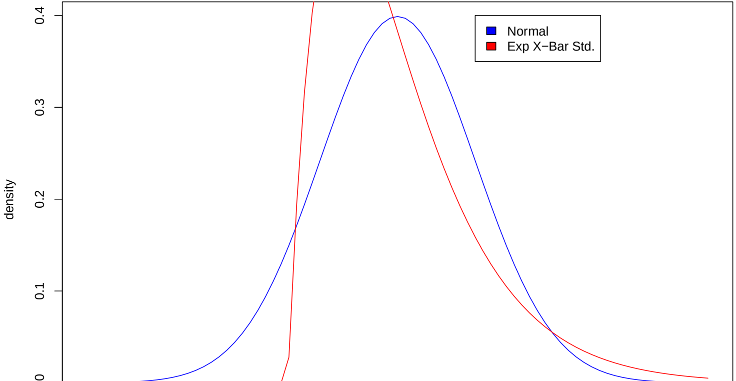
Now let the “underlying” distribution be Exp(1), which has mean and standard deviation both equal to $1/1 = 1$. Here’s the standardized density along with a $N(0,1)$ density:



how large is large enough case II - skewed

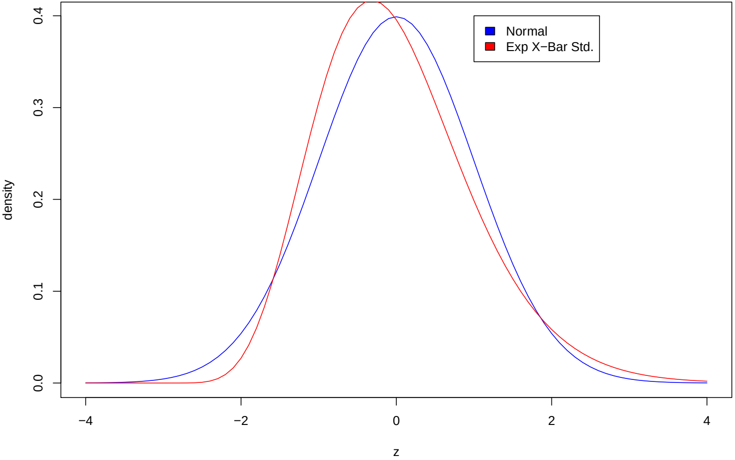
Now consider X_1, X_2 i.i.d. Exp(1), and its sample average $\overline{X}$, which will have mean 1 and standard deviation $1/\sqrt{2}$.

Here is a picture of the density for $\frac{\overline{X}-1}{1/\sqrt{2}}$, along with the density for $Z \sim N(0,1)$.



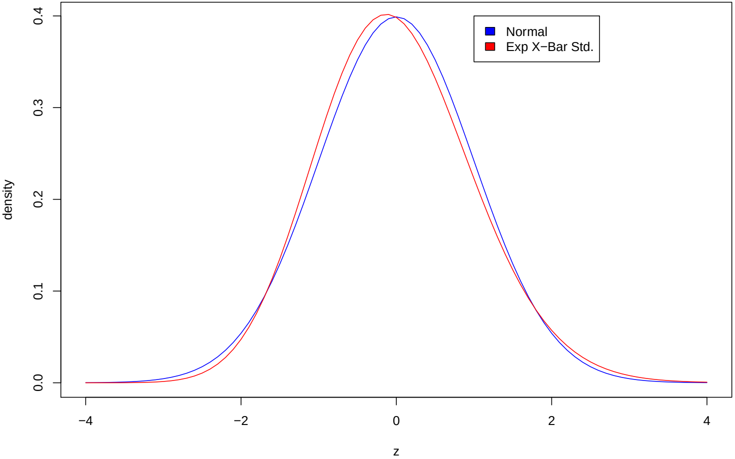
how large is large enough case II - skewed

Now with $n = 10$.



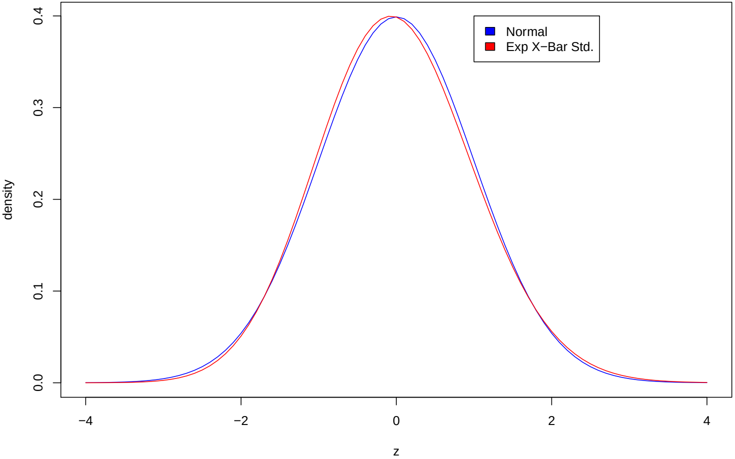
how large is large enough case II - skewed

Try $n = 60$



how large is large enough case II - skewed

Try $n = 200$



normal approximation applications

Given $X_1, \ldots, X_n$ i.i.d with mean μ and variance σ^2, as long as n is large enough, any of the following approximations hold. Pick the most convenient:

$$\sum_{i=1}^{n} X_i \sim^{approx} N(n\mu, \sqrt{n}\sigma)$$

$$\overline{X} \sim^{approx} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim^{approx} N(0, 1)$$

normal approximation example

Piston lifetimes in a Diesel engine follow a roughly symmetric distribution with mean 3.4 years and standard deviation 2.1 years.

What is the chance that the average life of 25 engines exceeds 4 years?

$$P(\overline{X} > 4) = P\left(\frac{\overline{X} - 3.4}{2.1/\sqrt{25}} > \frac{4 - 3.4}{2.1/\sqrt{25}}\right) \approx P(Z > 1.43) = 0.076358509537$$

another normal approximation example

A defective item is produced with probability $p = 0.01$. After $n = 10000$ items are produced, what is the probability that there were fewer than 80 defective items produced?

If X_i is 0 or 1 as item i is not defective, or defective, respectively, then $X_i \sim \text{Bernoulli}(0.01)$. We want $P\left(\sum_{i=1}^{n} X_i < 80\right)$.

In principle this is a $\text{Binomial}(n, p)$ calculation, but a very difficult one.

Using the normal approx. to sum of $\text{Bernoulli}(p)$ is straightforward:

$$\sum_{i=1}^{n} X_i \sim^{approx} N(n\mu, \sqrt{n}\sigma)$$

In this case $n = 10000$, $\mu = p$, and $\sigma^2 = p(1 - p)$. So:

$$\sum_{i=1}^{n} X_i \sim^{approx} N(100, \sqrt{99})$$

and

a quick note—n “large enough” in the Bernoulli(p) special case

The probability p is in some sense a “shape” parameter for Binomial distributions, in the sense that the close p is to 0 (or 1), the more right (left) skewed the distribution is.

It has been observed through experience and simulation that:

$$np \geqslant 5 \qquad \text{and} \qquad n(1-p) \geqslant 5$$

is sufficient for a good normal approximation.

Although this suggestion depends on p, which is usually unknown.

We’ll revisit this issue when the time comes to discuss ways to evaluate the empirical accuracy of a normal approximation.

more normal-based “sampling distributions”

The main focus will be on $\overline{X}$ and friends, because the most common statistical problem is to make statements about an unknown population mean μ and $\overline{X}$ will be used as the guess.

It turns out $\overline{X}$ is BFF with the sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

Why? Because the analysis of the properties of $\overline{X}$ will depend on:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{ (or } \sim^{approx}\text{) } N(0,1)$$

but typically σ is also unknown, and $\sqrt{S^2}$ will be used as *its* guess.

the distribution of the sample variance S^2 - I

First we'll consider the distribution of Z^2 when $Z \sim N(0,1)$.

$$
\begin{aligned}
M_{Z^2}(t) = E\left(e^{tZ^2}\right) &= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2(1-2t)}\, dz \\
&= \left(\frac{0.5}{0.5-t}\right)^{0.5} \int_{-\infty}^{\infty} \frac{1}{\left(\frac{0.5}{0.5-t}\right)^{0.5}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-0}{\left(\frac{0.5}{0.5-t}\right)^{0.5}}\right)^2} dz
\end{aligned}
$$

The integrand is a $N\left(0, \frac{0.5}{0.5-t}\right)$ density, so the integral equals 1, leaving

$M_{Z^2}(t) = \left(\frac{0.5}{0.5-t}\right)^{0.5}$

the distribution of the sample variance S^2 - II

A look at the list of mgfs reveals that $Z^2 \sim \text{Gamma}\left(\alpha = \frac{1}{2}, \lambda = \frac{1}{2}\right)$

Synonym: If $X \sim \text{Gamma}\left(\alpha = \frac{\nu}{2}, \lambda = \frac{1}{2}\right)$ then we give X a special name: "chi-square" distribution is parameter ν.

$$X \sim \chi^2_\nu$$

The parameter ν gets a curious name: *degrees of freedom*.

Theorem: If $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.

Therefore: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma)$, then:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n$$

Theorem: If X and Y are independent with $X \sim \chi^2_n$ and $X + Y \sim \chi^2_{n+m}$, then $Y \sim \chi^2_m$.

the distribution of the sample variance S^2 - III

Magical truth: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma)$, then $\overline{X}$ and S^2 are independent.

Finally: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma)$:

$$\begin{aligned}
\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X} + \overline{X} - \mu \right)^2 \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\overline{X} - \mu \right)^2 + 0 \\
&= \frac{n-1}{\sigma^2} S^2 + \left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \right)^2
\end{aligned}$$

Conclusion: $\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

the t distributions - I

I mentioned earlier that properties of $\overline{X}$ will depend on:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

and that σ will have to be guessed by $S = \sqrt{S^2}$.

However, when the constant σ is replaced with the random variable S, the result is no longer $N(0,1)$.

It turns out it is possible to derive the density of:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

the t distributions - II

The density $f(t)$ of T is is primarily nice to look at:

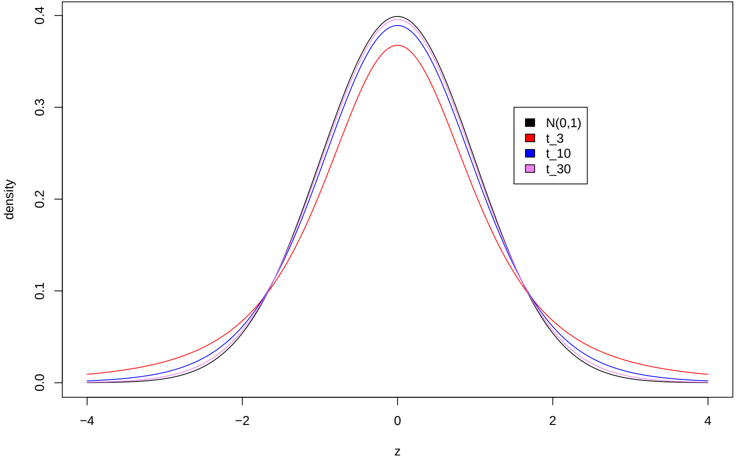
$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]\sqrt{\pi}\nu}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

where ν is called the "degrees of freedom" and in this case is $n-1$. You can think of $n-1$ as having been "inherited" from the denominator of T.

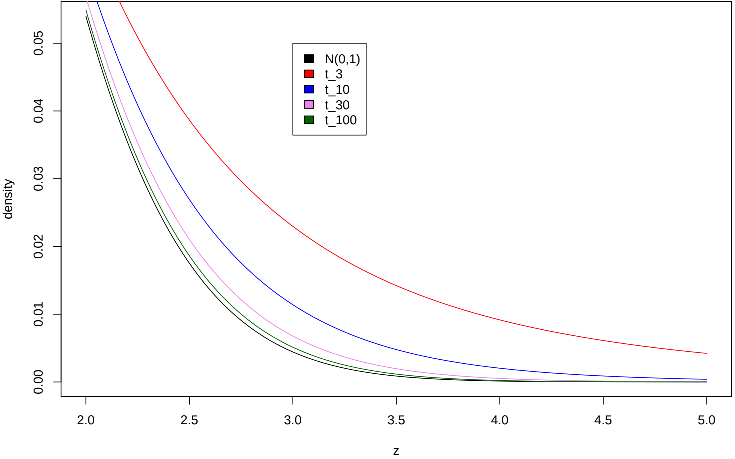
Important properties:

- Symmetric and bell shaped.
- As ν gets big, form of $f(t)$ approaches $e^{-t^2/2}\ldots$
- $\ldots$ in other words *as the sample size gets large, T starts to look like* $Z \sim N(0,1)$
- no anti-derivative, so a table of t probabilities needed on tests.

overall pictures of t_ν



pictures of t_ν in the "tail"



the F distributions

We might need the following technical result as well.

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ and $U \perp V$ then we say:

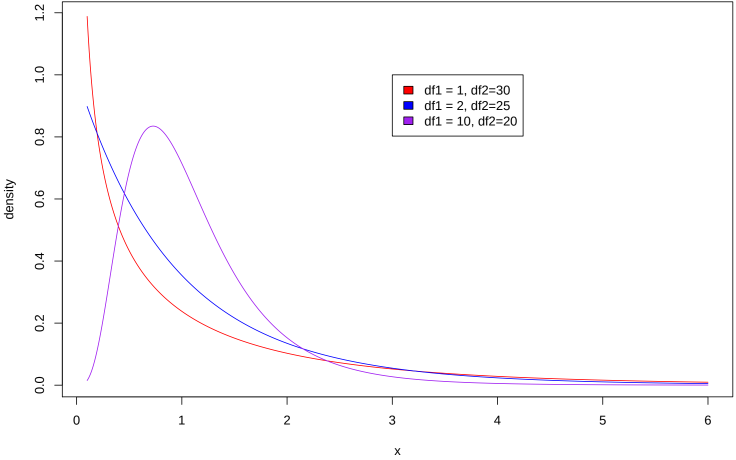
$$F = \frac{U/m}{V/n} \sim F_{m,n}$$

or "F has an F distribution with m and n degrees of freedom."

F distributions happen when it makes sense to consider the ratio of sums of squared normals. It is not yet obvious when this might take place.

The density is nasty, etc.

pictures of some F distributions



the t distributions - II-and-a-half

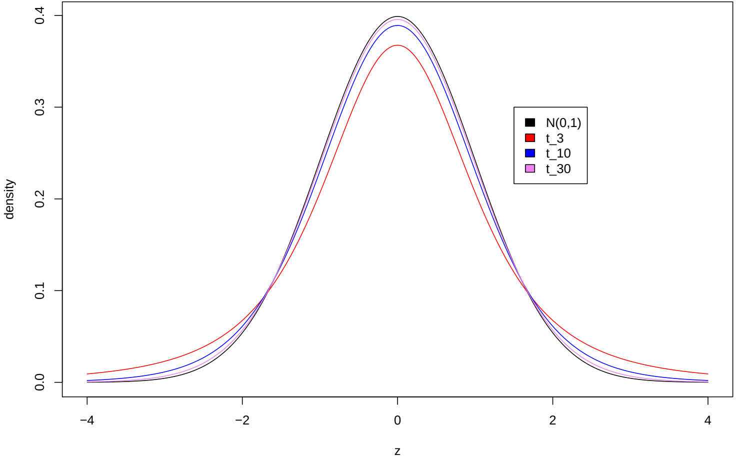
Rephrased from last time.

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t_{n-1} \qquad \left(\text{symbolically: } \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}\right)$$

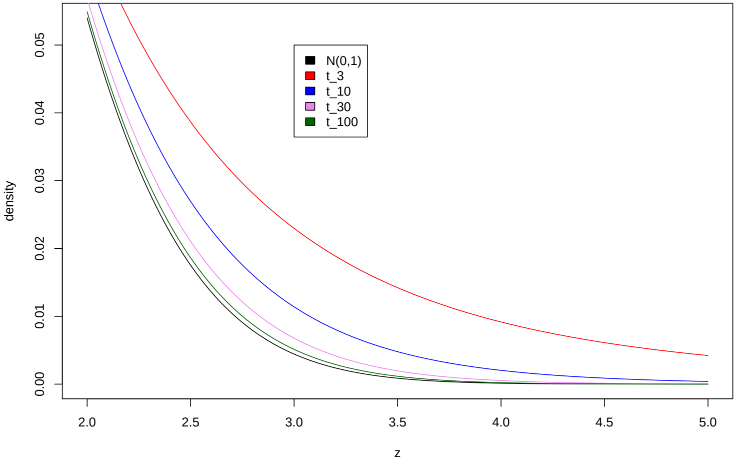
Things to remember:

- Symmetric and bell shaped.
- As n gets big, T starts to look like $Z \sim N(0,1)\}$
- a table of t probabilities needed on tests.

overall pictures of t_ν



pictures of t_ν in the "tail"



the F distributions

We might need the following technical result as well.

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ and $U \perp V$ then we say:

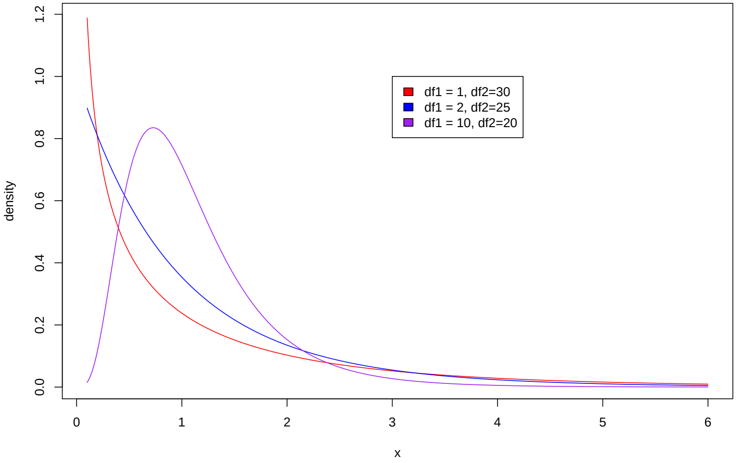
$$F = \frac{U/m}{V/n} \sim F_{m,n}$$

or "F has an F distribution with m and n degrees of freedom."

F distributions happen when it makes sense to consider the ratio of sums of squared normals. It is not yet obvious when this might take place.

The density is nasty; tables must be used on tests, etc.

pictures of some F distributions



preview example of F theory

Populations	Samples	Sample Variances
$N(\mu_1, \sigma_1)$	$X_{11}, \ldots, X_{1n_1}$	S_1^2
$N(\mu_2, \sigma_2)$	$X_{21}, \ldots, X_{2n_2}$	S_2^2

Question: are the two population standard deviations the same, or not?

Answer might be based on:

$$\frac{S_{n_1}^2/(n_1 - 1)}{S_{n_2}^2/(n_2 - 1)} \sim F_{n_1-1, n_2-1}$$

a specialized plot for detecting deviations from normality

central limit theorem, and friends

Recall that no matter the underlying distribution, as long as n is "large enough":

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim^{approx} N(0, 1)$$

But wait! There's more! As long as n is large enough (by exactly the same standard as before), we get something even more useful:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim^{approx} t_{n-1}$$

But the problem is. . . if we don't know the underlying distribution, how can we be confident in what sample size might be required.

The answer is to use *statistics*, by which I mean gather a sample and estimate what you don't know, which is in this case *how non-normal is the undelying distribution?*

limitations of "histogram"

During the first week of the course, the idea of histogram was used to motivate concepts of (empirical) symmetry and skewness.

But a histogram requires a very large sample size (hundreds?) to give an accurate picture. By the time you have n in the hundreds, the normal approximation is going to be pretty good.

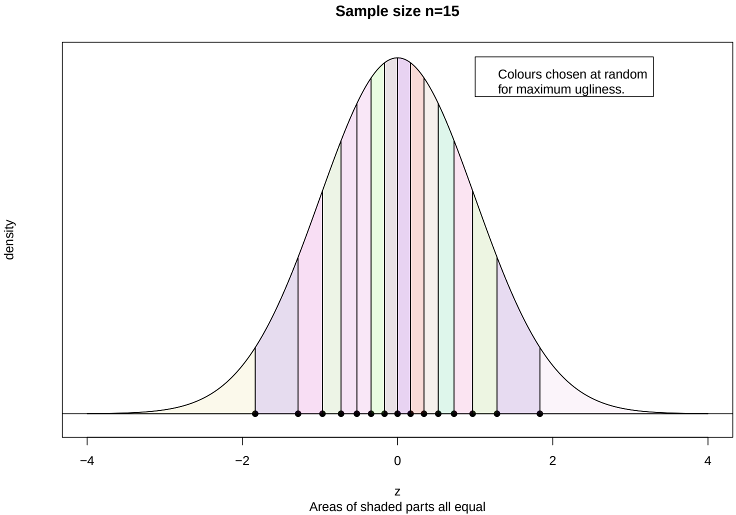
For the rest of the course we'll be concerned with distribution shape *relative to normality*, sometimes with samples too small for histograms.

There's an accurate graphical method that works as follows:

- (computer) puts the observed data in order.
- (computer) determines what a "perfect" $N(0, 1)$ dataset of the same sample size would look like.
- (computer) makes scatterplot with perfect (horizontal) vs. ordered data (vertical)
- straight line means data are consistent with having come from a normal distribution. Other patterns also easy to interpret.

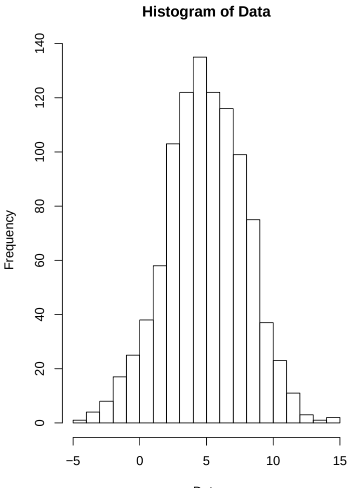
“perfect” standard normal data

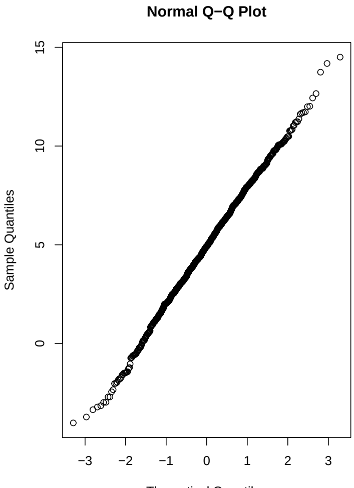
Find the values that split the area under the curve into equal parts.



result: “normal quantile plot” of “normal q-q plot” (other names)

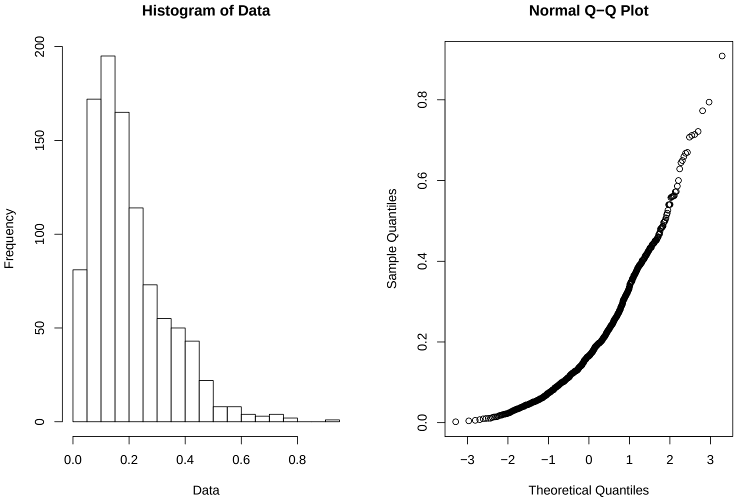
First set of examples will have $n = 1000$, in which case a histogram might have been OK anyway. First example: perfect normal $N(5, 3)$ data. Result: straight line.





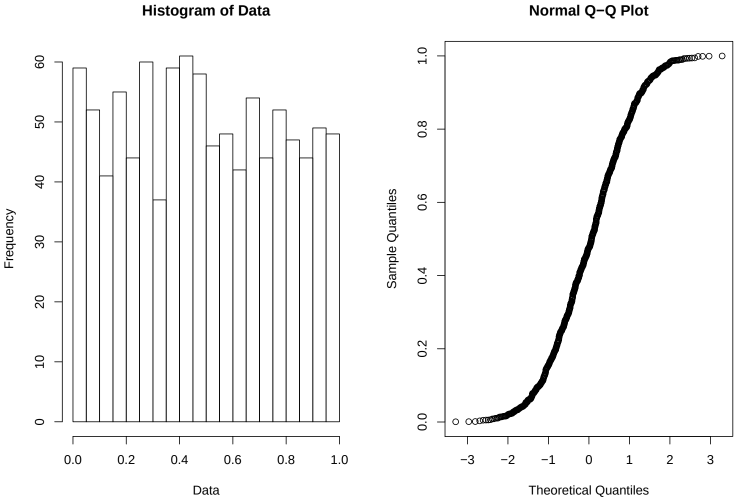
$n = 1000$ right skewed data

Gamma(2, 10). Result: curved (“concave ‘up’ ”)



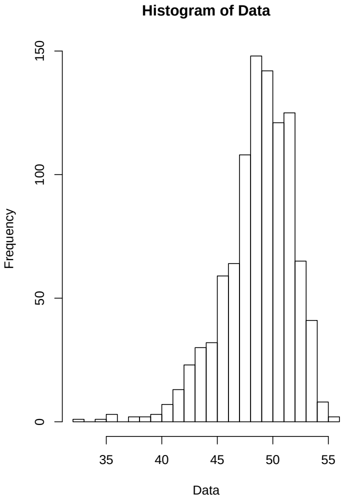
$n = 1000$ light tails

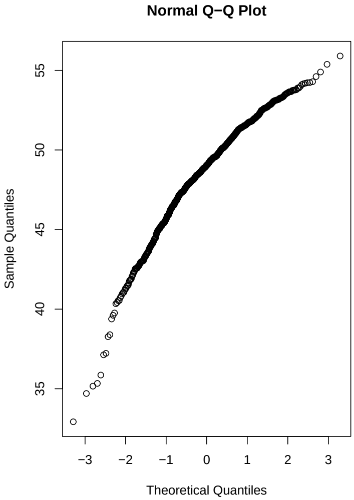
Uniform[0,1]. Result: S-shaped.



$n = 1000$ rarer case: left skewed

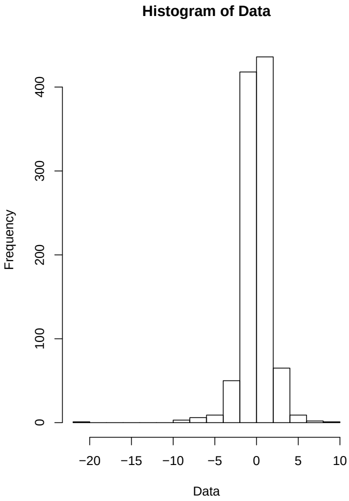
Result: curved (“concave ‘down’ ”)

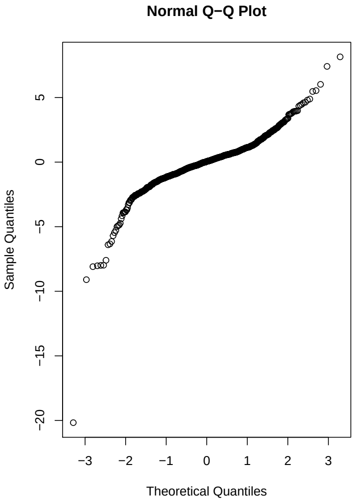




$n = 1000$ very rare: "heavy tails"

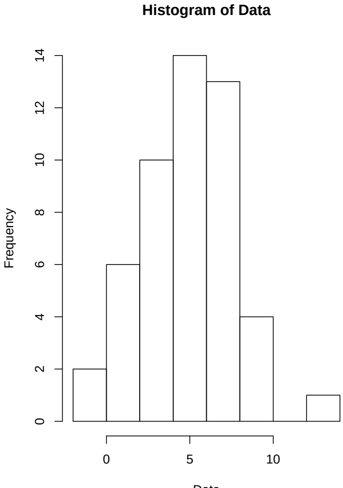
Result: "reverse-S" shaped

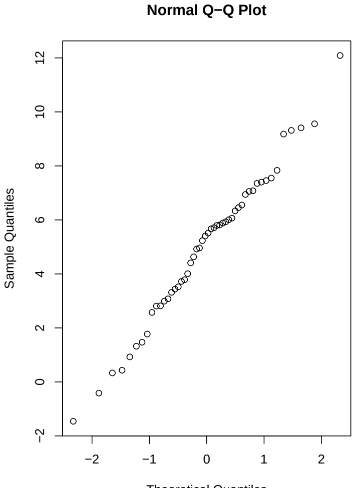




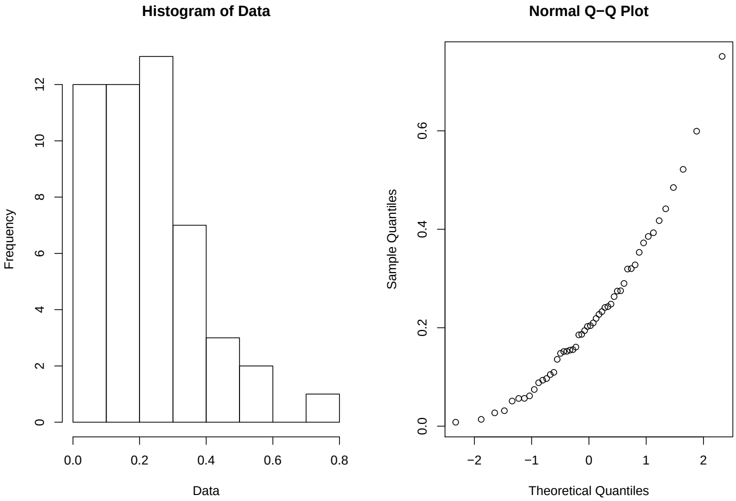
value of normal quantile plots is with small samples

Bring the sample size down to $n = 50$. Histogram not useful. Here's the $N(5, 3)$ example again.

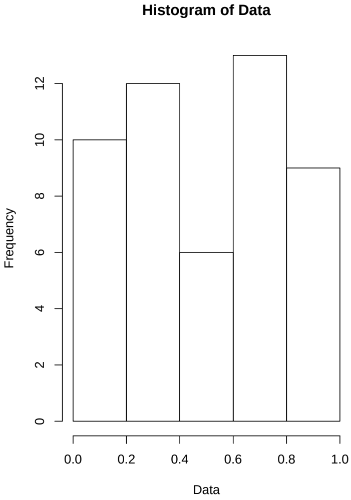




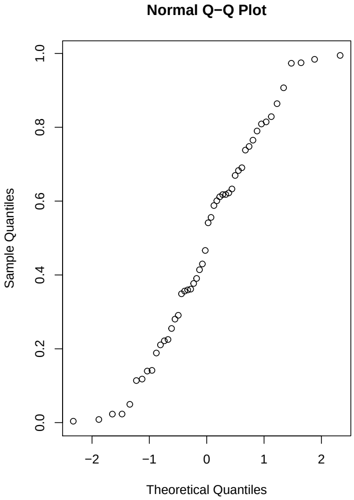
$n = 50$ right skewed



$n = 50$ light tails



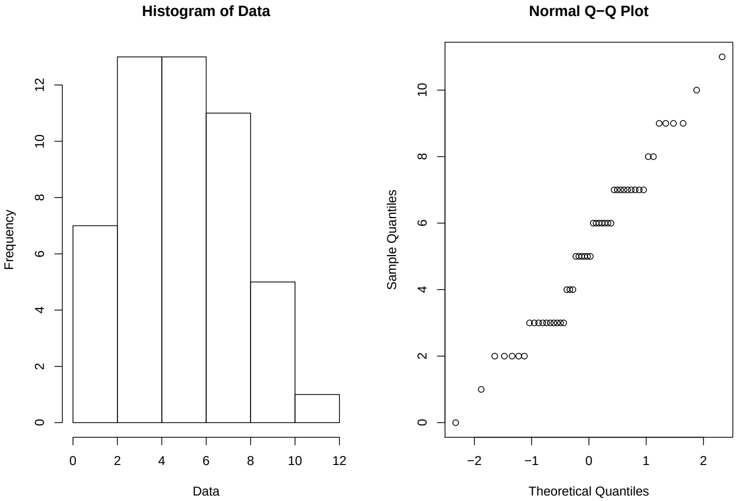
Histogram of Data



Normal Q−Q Plot

$n = 50$ from a discrete distribution

Don't worry about ties in the data. Just look at the overall pattern.



empirical verification of “sample size large enough”

Normal quantile plots are as good as you can get for determining deviations from normality, for the purpose of evaluating normal approximation accuracy.

There is one other important special case in which the underlying distribution is Bernoulli(p), which is not amenable to a normal quantile plot.

The rule when p is known was given as $np \geqslant 5$ and $n(1-p) \geqslant 5$.

When p is not known, but you have an observed sample $x_1, \ldots, x_n$, you can replace p in the rule with its empirical version:

$$\hat{p} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$$

roadmap

probability/statistics

I said probability was “done”, but it really wasn’t.

In statistics we use a sample to make statements about what is unknown about an underlying distribution.

Then we did more probability - but not for the purpose of modeling actual random processes in the wild.

The purpose of the additional probability was to determine some properties of functions of samples—with a focus on the properties of $\overline{X}$

Now we are actually going to do statistics.

parameter estimation

point estimation

We will treat population parameters as constants (as opposed to *Bayesian statistics*).

The goal is to use a statistic $\hat{\theta}$, i.e. a function of a sample, to estimate the value of a parameter θ, which could be a vector. This statistic is called a "point estimator".

e.g. Sample is from $N(\mu, 3)$. We want to estimate μ.

e.g. Sample is from Bernoulli(p). We want to estimate p.

e.g. Sample is from $N(\mu, \sigma)$. We want to estimate (μ, σ).

Open questions about point estimators:

- what desirable properties should they have?
- how do I know which one to use? (To be addressed later.)

bias and variance

We would like $\hat{\theta}$ to be *unbiased* ("for θ"):

$$E\left(\hat{\theta}\right) = \theta$$

For example, $\overline{X}$ is always unbiased for the mean μ of any population, because we showed earlier that:

$$E\left(\overline{X}\right) = \mu$$

There can be lots of unbiased estimators. How to choose the best one? Take the one with the smallest variance.

For example, if you have a sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma)$ distribution, then $\overline{X}$ is unbiased. But so is just taking X_1, say, because:

$$E(X_1) = \mu$$

But $\mathrm{Var}\left(\overline{X}\right) = \sigma^2/n$ which is smaller than $\mathrm{Var}(X_1) = \sigma^2$.

bias and variance

That was a silly example.

A less silly exampe is that is possible to show that the sample median $\tilde{X}$ is also unbiased for μ when the sample is from a normal population.

It turns out (FIXED - the problem was the 4 should have been 2) $\text{Var}\left(\tilde{X}\right) \approx \frac{\pi\sigma^2}{2n} \approx 1.57\text{Var}\left(\overline{X}\right)$, in the $N(\mu, \sigma)$ case.

So $\overline{X}$ is preferred.

Fact: when the population is normal, $\overline{X}$ is the unbiased estimator with the smallest variance.

Another desirable property (that $\overline{X}$ has, for example) is *consistency*, which means the variance tends to 0 as $n \to \infty$.

another unbiased estimator

Population: $N(\mu, \sigma)$. Sample: $X_1, \ldots, X_n$.

Paramater to estimate: σ^2.

Since:

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

and the expected value of a Gamma(α, λ) is $\frac{\alpha}{\lambda}$, we get:

$$E\left(\frac{n-1}{\sigma^2} S^2\right) = n - 1$$

and therefore $E(S^2) = \sigma^2$.

This explains the embarassing $n-1$ in the denominator of S^2.

$Exp(\lambda)$

How to estimate the rate parameter λ?

interval estimation

estimation, with an assessment of the data collection plan

We'll come back to the problem of determining a good estimator from first principles.

A shortcoming of a *point estimator* is that is doesn't suggest how far wrong it might be.

A better option is to provide a pair of estimators $\hat{\theta}_L$ and $\hat{\theta}_U$ that satisfy the following equation:

$$P\left(\hat{\theta}_L < \theta < \hat{\theta}_U\right) = 1 - \alpha$$

for some pre-determined α.

α can be anything between 0 and 1, but is typically chosen to be small.

α is arbitrary. There is no "correct" or "better" α value. By far and away the most common choice is $\alpha = 0.05$.

The interval $\left[\hat{\theta}_L, \hat{\theta}_U\right]$ is called a $(1 - \alpha) \cdot 100\%$ *confidence inteval* for θ.

It is possible to have $\hat{\theta}_L = -\infty$ or $\hat{\theta}_U = \infty$

When $\alpha = 0.05$ (as usual), we have a 95% confidence interval.

(artifical) example of a confidence interval

Suppose the underlying population is $N(\mu, \sigma_0)$ with σ_0 (magically) known.

We plan to gather a sample $X_1, \ldots, X_n$. There are *lots* of 95% confidence intervals for μ, obtained by isolating μ in the middle of:

$$P\left(a < \frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} < b\right) = 1 - \alpha$$

Define z_α as the solution of $P(Z \leqslant z_\alpha) = 1 - \alpha$, where $Z \sim N(0,1)$. The *shortest possible* 95% confidence interval for μ comes from:

$$P\left(\overline{X} - z_{0.025}\frac{\sigma_0}{\sqrt{n}} < \mu < \overline{X} + z_{0.025}\frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

a near-universal 95% C.I. formula

Note that $z_{0.025} = 1.96$.

Also, the standard deviation of $\overline{X}$ is $\frac{\sigma_0}{\sqrt{n}}$

A synonym for the phrase "standard deviation of the estimator" is *standard error*, abbreviated: s.e.

Neil's patented universal 95% C.I. formula is:

$$\text{estimator} \pm \text{“2”s.e.(estimator)}$$

"Two" is in quotation marks because the precise value will vary a little over and under 2, but it will always be close to 2 (for a 95% interval).

meaning, and some myths

A confidence interval is a statement about the “plan” to gather a sample.

C.I. meaning: The “plan” will result in an interval that will capture the true parameter with probability $1 - \alpha$.

Once the dataset is collected, and the numbers plugged into the suitable C.I. formula, that is a realization of the C.I. formula.

Most common myth goes like this. The dataset is collected, and the 95% confidence interval for μ is, say [4.2, 6.8].

> *There is a 95% chance that μ is between 4.2 and 6.8.*

The statement is nonsense. Either μ is between 4.2 and 6.8, or it isn't.

things that affect the width of a typical C.I.

The (artificial) example is nevertheless characteristic:

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

The larger the σ_0, the wider the C.I. (But you have no control over σ_0.)

The larger the α, the narrower the C.I. (But α is arbitrary.)

The larger the n, the narrower the C.I. (The sample size *is* under your control.)

a sample size calculation

The *(absolute) margin of error* e is half the width of a confidence interval.

In the current (artificial) situation, to produce a $(1-\alpha)\cdot 100\%$ confidence interval of width $2e$, the sample size needs to be:

$$n = \left(\frac{z_{\alpha/2}\sigma_0}{e}\right)^2$$

This won't usually be an interger, so to drive students crazy I say "pick one of the two sample sizes adjacent", because it really doesn't matter.

In reality σ is not known. There are a few practical options:

- collect a "pilot sample" of some moderate size (30 to 50, say), to get an estimate of σ.
- use prior knowledge of the value of σ
- if the population is plausibly normal, use prior knowledge of the minimum m and maximum M plausible values you might ever see, and use $(M-m)/6$ as a rough guesstimate for σ.

the classic "one-sample" t interval - I

A more realistic situation is that the population is $N(\mu, \sigma)$, both parameters unknown, although the mean is of primary interest. We plan to get a sample $X_1, \ldots, X_n$.

The (artificial) interval was based on:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

We'll do this often—replace σ with S—to obtain:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Define $t_{n-1,\alpha}$ as the solution of $P(t_{n-1} \leqslant t_{n-1,\alpha}) = 1 - \alpha$. The new interval will be based on:

$$P\left(-t_{n-1,\alpha/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1 - \alpha$$

the classic "one-sample" t interval - II

The interval is therefore:

$$\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

The value $\frac{S}{\sqrt{n}}$ is (also) called the (estimated) standard error for $\overline{X}$, or s.e.$(\overline{X})$, and in the usual 95% case we end up with another example of the universal formula:

$$\text{estimator} \pm \text{"2"s.e.(estimator)}$$

(patent pending) for any non-insane sample size.

That's because t calculations aren't wildly different from $N(0,1)$ calculations as long as $n-1$ isn't tiny.

$t_{n-1,0.025}$ for some non-insane sample sizes

n	t
15	2.13144954556
30	2.04227245630
40	2.02107539031
50	2.00855911210
60	2.00029782201
150	1.97590533090

example

Examples themselves tend to be as interesting as watching paint dry.

So, for example, consider textbook question 9.14, which gives 15 values for the drying time, in hours, of a brand of latex paint.

x_bar	s	n
3.78666666667	0.970910225853	15

The 95% confidence interval for the mean drying time is:

conf.low	conf.high
3.24899450507	4.32433882826

watching even more paint dry

The paint company wants to estimate the mean paint drying time to within a margin of error of 10 minutes, with 95% confidence.

Sample size requirement: The formula is:

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

We could use our best guess from the available information $s = 0.971$ (in hours) in place of σ in the calculation:

$$\left(\frac{1.96 \cdot 0.971}{10/60}\right)^2 = 130.369$$

So just to bother you, I'll use $n = 130$.

gather a sample of size $n = 130$

Here is a relevant summary of the dataset:

x_bar	s	n
4.16	1.02	130

From the t_{129} distribution we get $t_{129,0.025} = 1.979$. So the 95% confidence interval is:

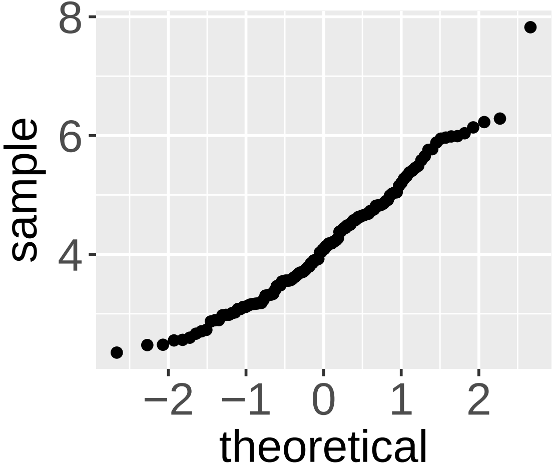
$$\overline{x} \pm t_{129,0.025}\frac{s}{\sqrt{n}} = 4.157 \pm 1.979\frac{1.022}{\sqrt{130}}$$

or

$$[3.98, 4.334]$$

verifying the model assumption(s)

In this case there is only one assumption (that can be verified)—that the underlying distribution is normal. Here is a normal quantile plot of the data:



model assumption conclusion | robustness of "t procedure"

The normal distribution assumption has been violated. It seems the underlying distribution is skewed right.

However, the sample size $n = 130$ is large, so by the speed of convergence of the CLT and its buddy Mr. Slutsky, we're still OK.

Dirty secret: I simulated the data to get the example sample of size 130.

As an example of what is called the "robustness" of this confidence interval against violations of the normality assumption, I did a quick simulation (code embedded in notes).

The proportion of the 10^4 simulated confidence intervals that captured the true mean is (for this simulation—changes every time I render the lecture notes):

$$0.9457$$

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\text{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

The variance of this expression is:

$$\text{Var}\left(\overline{X} - X\right) = \text{Var}\left(\overline{X}\right) + \text{Var}(X) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

If the population is normal, so will be $\overline{X} - X$, and its mean will be $E\left(\overline{X} - X\right) = \mu - \mu = 0$

prediction “interval”

Put it all together to get:

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0,1)$$

Deal with unknown σ right now—replace it with S from the sample, to get:

$$\frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

A $100 \cdot (1 - \alpha)\%$ *prediction interval* can be obtained by solving for X in:

$$P\left(-t_{n-1,\alpha/2} < \frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} < t_{n-1,\alpha/2}\right)$$

prediction interval example

The formula is:

$$\overline{X} \pm t_{n-1,\alpha/2} S \sqrt{1 + \frac{1}{n}}$$

Just that little ‘1’ under the square root—but it makes all the difference. It guarantees that the prediction can never be better than the variance in the population itself, no matter what the sample size.

Using the paint example, with $t_{129,0.025} = 1.97852$ and

x_bar	s	n
4.16	1.02	130

in the formula gives:

$$4.157 \pm 1.979 \cdot 1.022 \sqrt{1 + \frac{1}{130}} \qquad \text{or} \qquad [2.128, 6.186]$$

prediction interval model assumptions

Normal population is the only assumption.

Suppose the population is not normal. What might happen to the following as n gets large?

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}}$$

There is no way of knowing. X just sits there in the numerator, with properties that never change no matter what the sample size.

So the population really has to be normal, or the P.I. formula doesn't work.

The paint drying P.I. we calculated is therefore not that useful.

the two-sample problem (normal populations)

We've solved the case of one numerical variable in a dataset with a normal population.

Often you'll have a numerical variable in one column, and a "grouping" variable in another column that categorizes the observations into two groups.

Variable	Group
3.85	2
6.06	2
3.28	1
4.85	2
5.34	1
6.03	2
$\vdots$	$\vdots$

Variable	Group
X_{21}	2
X_{22}	2
X_{11}	1
X_{23}	2
X_{12}	1
X_{24}	2
$\vdots$	$\vdots$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, with the following properties:

$$\begin{aligned} E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\ \mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right) &= \mathrm{Var}\left(\overline{X}_1\right) + \mathrm{Var}\left(\overline{X}_2\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

We need to figure out what to do about σ^2.

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The “obvious” estimator is $\overline{X}_1 - \overline{X}_2$, which will have a normal distribution with:

$$\begin{aligned} E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\ \text{Var}\left(\overline{X}_1 - \overline{X}_2\right) &= \text{Var}\left(\overline{X}_1\right) + \text{Var}\left(\overline{X}_2\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

We need to figure out what to do about σ^2.

for σ^2, use the data from both samples

The sample variances S_1^2 and S_2^2 are both unbiased estimators for σ^2, so any weighted average (with weights that add up to 1) of them will also be an unbiased estimator.

We call the following sample-size-based choice of weights the *pooled sample variance* estimator for σ^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2}{(n_1 - 1) + (n_2 - 1)} + \frac{(n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Note that when $n_1 = n_2$ this is just the average of the two sample variances.

putting it all together

We want an interval estimator for $\mu_1 - \mu_2$. So far we have:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

Who wants to guess what the distribution of this will be:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Isolating $\mu_1 - \mu_2$ in the usual way gives the confidence interval formula.

two normal samples, equal variances C.I.

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{n_1+n_2-2,\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In the 95% case, another instance of my patented:

$$\text{estimator} \pm \text{“2”s.e.(estimator)}$$

example - watching two kinds of paint dry

If the world of one brand of paint drying was too fast-paced, this example is for you. (Question 9.49 from the textbook.) Two brands of paint will have their drying times compared.

The goal is to estimate the difference between the mean drying times.

Here's a glance at the "dataset" as is, organized in a way it should never be collected:

```
## # A tibble: 15 × 2
##    PaintA PaintB
##     <dbl>  <dbl>
## 1     3.5    4.7
## 2     2.7    3.9
## 3     3.9    4.5
## 4     4.2    5.5
## 5     3.6    4.0
## 6     2.7    5.3
## 7     3.3    4.3
```

watching two kinds of paint dry

A real dataset looks like this:

```
## # A tibble: 30 × 2
##       brand  time
##       <chr> <dbl>
## 1  PaintA   3.5
## 2  PaintA   2.7
## 3  PaintA   3.9
## 4  PaintA   4.2
## 5  PaintA   3.6
## 6  PaintA   2.7
## 7  PaintA   3.3
## 8  PaintA   5.2
## 9  PaintA   4.2
## 10 PaintA   2.9
## # ... with 20 more rows
```

watching two kinds of paint dry

Anyway, here is a summary of the two groups:

brand	x_bar	samp_var	n
PaintA	3.82	0.607	15
PaintB	4.94	0.568	15

The degrees of freedom is 28. The number from the t distribution is $t_{28,0.025} = 2.048$.

The 95% confidence interval is $[-1.693, -0.547]$.

when the variances cannot be assumed to be equal

Most two-sample analyses in practice don't bother with the equal variance assumption, and just use the following sequence of facts.

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim^{approx} t_\nu$$

where ν has one of the most disgraceful formulae in the history of formulae. Don't look at its formula on the next slide.

what did I just tell you?

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

This formula is not for humans to use.

But there are few things to notice about it:

- if $S_1^2 \approx S_2^2$, then $\nu \approx n_1 + n_2 - 2$.
- if $S_1^2 \ll S_2^2$, then $\nu \approx n_2 - 1$, and vice versa.

It won't usually be an integer, so if you need to use this method on a test (where I'd give you the value of ν), just use whatever nearby integer that is convenient.

watching two kinds of paint dry, now in a very slightly different way

The C.I. formula becomes:

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

which in the 95% case is another instance of my patented formula.

The paint drying example, redux:

brand	x_bar	samp_var	n
PaintA	3.82	0.607	15
PaintB	4.94	0.568	15

The degrees of freedom is $\nu = 27.969$. The number from the t distribution is $t_{27.969,0.025} = 2.049$.

The 95% confidence interval is $[-1.694, -0.546]$ (as compared to $[-1.693, -0.547]$)

watching plants grow

Instead of watching paint dry, let's watch plants grow. (Textbook question 9.40.)

20 tree seeds are planted. 10 get a nitrogen fertilizer. After 140 days all the stem growths are measured in grams. The goal is to estimate the mean difference between the two groups.

Here is a summary of the data. It is unlikely that the group variances are equal.

fertilizer	x_bar	samp_var	n
Nitrogen	0.565	0.035	10
NoNitrogen	0.399	0.005	10

The degrees of freedom is $\nu = 11.673$. The number from the t distribution is $t_{11.673,0.025} = 2.186$.

The 95% confidence interval is $[0.027, 0.305]$.

two-sample t procedure model assumptions

There are two assumptions.

1. The populations are both normal.
2. The samples are independent.

The normality assumption is assessed by looking at normal quantile plots of both samples. If there is a violation, a large sample size for that sample makes the problem go away.

The independence assumption can only be assessed by considering the way the dataset was collected.

There is one very common way to collect samples that are not independent, and that is to collect two observations, say X_{i1} and X_{i2} on the i^{th} experimental unit. We will examine this situation next.

the patented four-step procedure outline for confidence intervals

The goal is to get $P\left(\hat{\theta}_L < \theta < \hat{\theta}_U\right) = 1 - \alpha$

1. Decide what θ is.
2. Decide what $\hat{\theta}$ is (depends on how dataset was collected.)
3. Compute $\mathrm{Var}\left(\hat{\theta}\right)$.
4. Deal with unknowns in part 3.,and settle for $\widehat{\mathrm{Var}}\left(\hat{\theta}\right)$

We've handled the one- and two-normal sample cases.

one normal sample masquerading as two - I

In the two-sample case, the samples are assumed to be independent, which was crucial for the calculation of $\text{Var}\left(\hat{\theta}\right)$.

It is possible (common, even) to collect a sample $X_{11}, \ldots, X_{1n}$ from $N(\mu_1, \sigma_1)$ and another sample $X_{21}, \ldots, X_{2n}$ from $N(\mu_2, \sigma_2)$, where X_{1i} and X_{2i} are measured on the same (i^{th}) "experimental unit".

We might still be interested in the parameter $\theta = \mu_D = \mu_1 - \mu_2$.

We will still use the "obvious" $\hat{\theta} = \overline{X}_1 - \overline{X}_2$, in a sense.

But the better way to express $\hat{\theta}$ is to consider the differences $D_i = X_{1i} - X_{2i}$ and use $\overline{D}$.

We still have $\overline{D}$ normal with mean $E(D_i) = \mu_1 - \mu_2$. But the variance of $\overline{D}$ will be $\text{Var}(D_i)/n$, where:

$$\text{Var}(D_i) = \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\text{Cov}(X_{1i}, X_{2i})$$

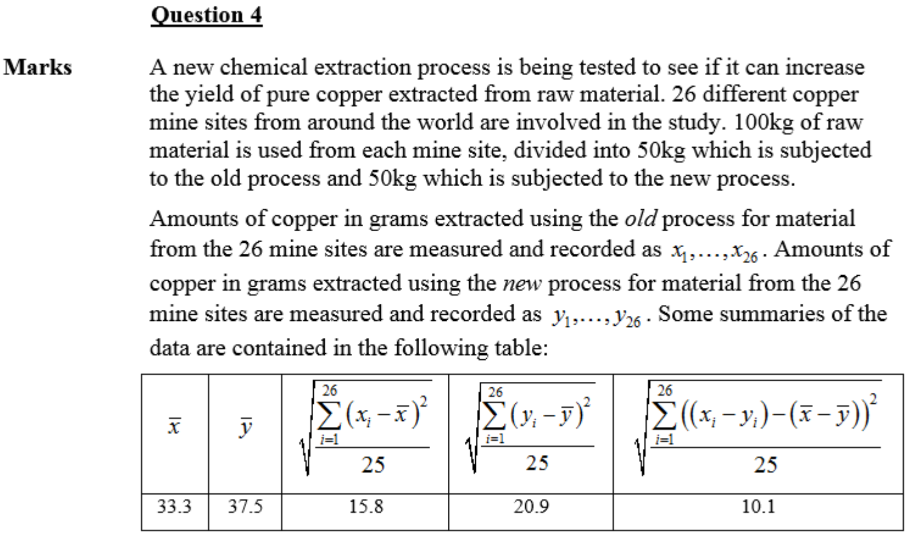
one normal sample masquerading as two - II

We treat the “two” samples for what they are, which is really one sample $D_1, D_2, \ldots, D_n$ i.i.d. $N(\mu_1 - \mu_2, \sigma_D)$.

And we already know how to analyze the one independent sample case.

The only challenge seems to be do determine when there are two independent samples, or only one sample of differences.

example - 2007 exam



Question 4

Marks

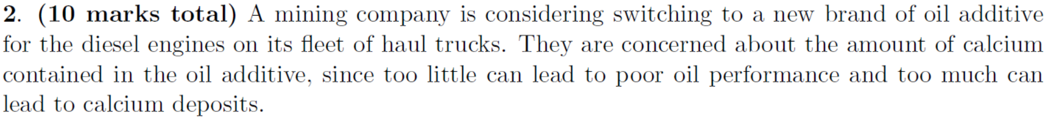
A new chemical extraction process is being tested to see if it can increase the yield of pure copper extracted from raw material. 26 different copper mine sites from around the world are involved in the study. 100kg of raw material is used from each mine site, divided into 50kg which is subjected to the old process and 50kg which is subjected to the new process.

Amounts of copper in grams extracted using the *old* process for material from the 26 mine sites are measured and recorded as $x_1, \ldots, x_{26}$. Amounts of copper in grams extracted using the *new* process for material from the 26 mine sites are measured and recorded as $y_1, \ldots, y_{26}$. Some summaries of the data are contained in the following table:

$\bar{x}$	$\bar{y}$	$\sqrt{\dfrac{\sum_{i=1}^{26}(x_i-\bar{x})^2}{25}}$	$\sqrt{\dfrac{\sum_{i=1}^{26}(y_i-\bar{y})^2}{25}}$	$\sqrt{\dfrac{\sum_{i=1}^{26}\left((x_i-y_i)-(\bar{x}-\bar{y})\right)^2}{25}}$
33.3	37.5	15.8	20.9	10.1

Figure 4:

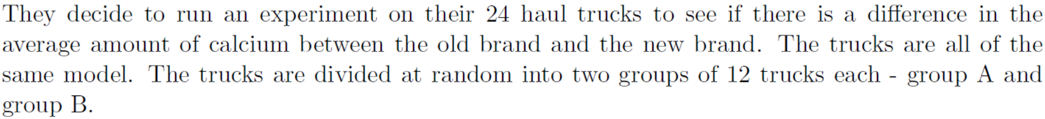
example - 2012 MIE237 exam - I



2. (10 marks total) A mining company is considering switching to a new brand of oil additive for the diesel engines on its fleet of haul trucks. They are concerned about the amount of calcium contained in the oil additive, since too little can lead to poor oil performance and too much can lead to calcium deposits.

Figure 5:

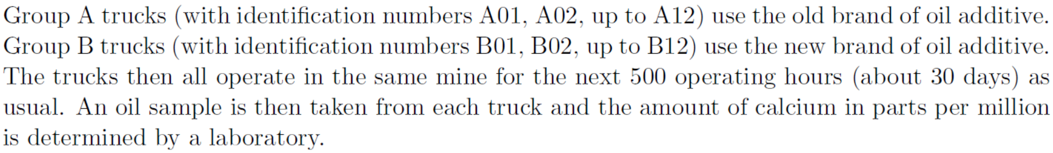
example - 2012 MIE237 exam - II



They decide to run an experiment on their 24 haul trucks to see if there is a difference in the average amount of calcium between the old brand and the new brand. The trucks are all of the same model. The trucks are divided at random into two groups of 12 trucks each - group A and group B.

Figure 6:

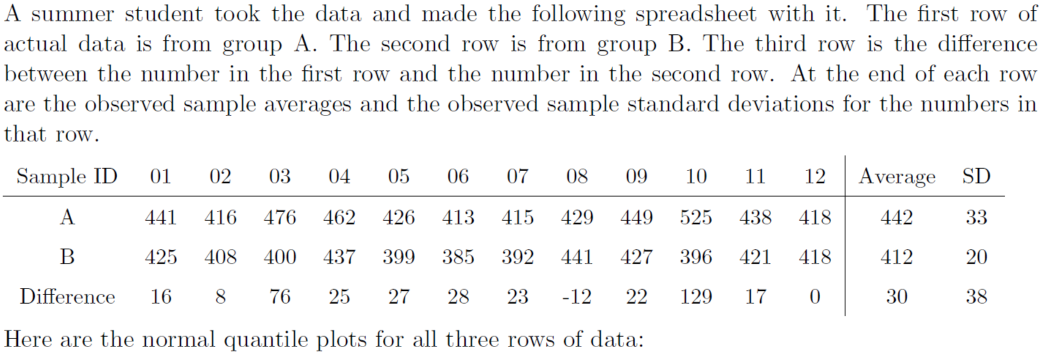
example - 2012 MIE237 exam - III



Group A trucks (with identification numbers A01, A02, up to A12) use the old brand of oil additive. Group B trucks (with identification numbers B01, B02, up to B12) use the new brand of oil additive. The trucks then all operate in the same mine for the next 500 operating hours (about 30 days) as usual. An oil sample is then taken from each truck and the amount of calcium in parts per million is determined by a laboratory.

Figure 7:

example - 2012 MIE237 exam - IV



A summer student took the data and made the following spreadsheet with it. The first row of actual data is from group A. The second row is from group B. The third row is the difference between the number in the first row and the number in the second row. At the end of each row are the observed sample averages and the observed sample standard deviations for the numbers in that row.

Sample ID	01	02	03	04	05	06	07	08	09	10	11	12	Average	SD
A	441	416	476	462	426	413	415	429	449	525	438	418	442	33
B	425	408	400	437	399	385	392	441	427	396	421	418	412	20
Difference	16	8	76	25	27	28	23	-12	22	129	17	0	30	38

Here are the normal quantile plots for all three rows of data:

Figure 8:

when the parameter is a single probability

We have solved the problem of estimating μ from a single normal sample.

It's common to have a variable in the dataset that only takes on two values, which we would model using a Bernoulli(p) distribution (effectively treating the values as 0's and 1's.)

The sample is the sequence of random 0's and 1's: $X_1, X_2, \ldots, X_n$ i.i.d. Bernoulli(p).

Patented 4-step procedure:

1. $\theta = p$
2. The "obvious" estimator is $\overline{X}$, which we will more commonly call in this special case the "sample proportion" or $\hat{p}$. This is an unbiased estimator for p.
3. The variance is $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$, which also contains p.
4. To deal with the unknown p, we just use the sample itself and plug in $\hat{p}$ (!)

when the parameter is a single probability

As long as $n\hat{p}$ and $n(1-\hat{p})$ both exceed 5, we know the following is a good approximation:

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

This gives us our desired expression:

$$P\left(-1.96 < \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < 1.96\right) \approx 0.95$$

giving us a 95% confidence interval for the unknown probability:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

which is yet another example of my patented C.I. formula.

example

Work Team Alpha inspects 8000 gas meters. They find 87 defective ones. Produce a 95% confidence interval for the probability that a gas meter is defective.

comparing two proportions

One variable in the dataset with 0's and 1's; another variable splitting observations into two groups.

The two populations are Bernoulli(p_1) and Bernoulli(p_2). The independent samples are $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$

Patented process:

1. Estimate $\theta = p_1 - p_2$.
2. Estimator $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{X_1} - \overline{X_2}$
3. $\text{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
4. Whoops! Don't know p_1 or p_2. So use $\hat{p}_1$ and $\hat{p}_2$ instead. Bam. Done.

Formula for 95% interval:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

two non-robust confidence intervals

Every procedure I have explained so far is *robust* as long as the sample size is large enough (except for the prediction interval formula.)

In principle we could apply the patented procedure to estimate σ^2 with S^2, using a χ^2 distribution.

We could also apply the patented procedure to estimate the ratio σ_1^2/σ_2^2 with S_1^2/S_2^2 using an F distribution.

But the results are well known to be non-robust, even with large sample sizes, so I cannot recommend them for use.

chiselling your own estimators onto stone tablets

fun fact from mathematics

Suppose a twice-differentiable function $f(x)$ has a critical value at x_0, and $g(x)$ is strictly increasing and twice-differentiable.

Then $g(f(x))$ also has a critical value at x_0, and the sign of its second derivative at x_0 is the same as the sign of the second derivative of f at x_0.

This can be seen by evaluating the left hand sides at x_0:

$$\begin{aligned}(g(f(x)))' &= g'(f(x))f'(x) \\ (g(f(x)))'' &= g''(f(x))(f'(x))^2 + g'(f(x))f''(x)\end{aligned}$$

estimating a proportion, from first principles - I

Here's a simulated sequence of 0's and 1's from a Bernoulli(p) distribution. I know what (p), but you don't.

```
##  [1] 0 0 1 1 0 0 1 0 0 0
```

What value of p between 0 and 1 is the *most likely* to have produce this sequence of 3 1's and 7 0's?

The probability of getting this sample exactly is:

$$(1-p)\cdot(1-p)\cdot p\cdot p\cdot(1-p)\cdot(1-p)\cdot p\cdot(1-p)\cdot(1-p)\cdot(1-p)$$
$$=p^3(1-p)^7$$

Let's call this function $L(p)$.

We could maximize $L(p)$, but it's easier to maximize $\ell(p)=\log L(p)$

estimating a proportion, from first principles - II

$$0 = \frac{d}{dp}\ell(p) = \frac{d}{dp}\left(3\log(p) + 7\log(1-p)\right) = \frac{3}{p} - \frac{7}{1-p}$$
$$\frac{3}{p} = \frac{7}{1-p} \Longrightarrow p = \frac{3}{10}$$

The second derivative is negative, so this is a maximum.

It would have been no harder to work in general, with k 1's out of a sample of size n, and maximizing $L(p) = p^k(1-p)^{n-k}$

The same calculus gives the maximum at k/n.

This is exactly the same as $\hat{p}$ that was used as "obvious" from before.

“likelihood function” for Bernoulli

The p.m.f. of a Bernoulli(p) is $f(x;p) = p^x(1-p)^{1-x}$ with $x \in \{0,1\}$.

Given a sequence $\{x_1,\ldots,x_n\}$ of 0’s and 1’s, yet another way of constructing $L(p)$ is as follows:

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ &= \prod_{i=1}^{n} f(x_i;p) \end{aligned}$$

Also: $\ell(p) = \log L(p) = \sum_{i=1}^{n} \log f(x_i;p)$

likelihood function in general

Given a sequence of observations $\{x_1, \ldots, x_n\}$ ("the data") from a random variable X with pmf or pdf $f(x;\theta)$, a likelihood function $L(\theta) = L(x_1, \ldots, x_n; \theta)$ for the parameter θ is defined as (for any positive g):

$$L(\theta) = \underbrace{g(\mathbf{x}) \prod_{i=1}^{n} f(x_i;\theta)}_{\text{real definition}} \propto \underbrace{\prod_{i=1}^{n} f(x_i;\theta)}_{\text{easy definition}}$$

If X is discrete and f is a pmf, then $L(\theta)$ is literally the probability of the data given θ.

If X is continuous and f is a pdf, then $L(\theta)$ is not a probability, but it still provides a useful "index" for θ values.

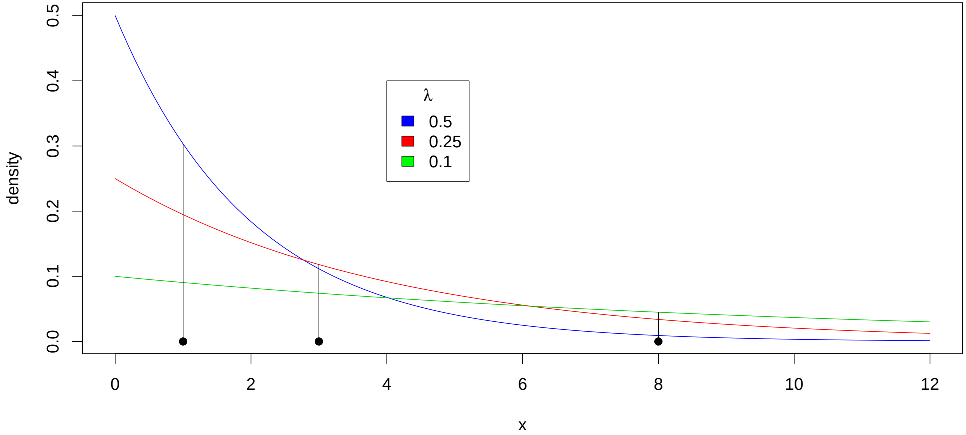
likelihood as “index” in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

a possibly useless and confusing picture



likelihood as “index” in continuous case

Suppose $X \sim \text{Exp}(\lambda)$ and the data are: $1, 3, 8$. A likelihood for λ is:

$$L(1,3,8;\lambda) = \lambda^3 e^{-\lambda(1+3+8)} = \lambda^3 e^{-12\lambda}$$

Consider three possible candidate guesses for the true value of λ: 0.1, 0.25, and 0.5.

$$\begin{aligned} L(0.5) &= 0.5^3 e^{-12} = 3.098 \times 10^{-4} \\ L(0.25) &= (0.25)^3 e^{-12 \cdot 0.25} = 7.779 \times 10^{-4} \longleftarrow \text{ Highest "likelihood"} \\ L(0.1) &= (0.1)^3 e^{-12 \cdot 0.1} = 3.012 \times 10^{-4} \end{aligned}$$

maximum likelihood “estimate”

The value of θ that maximizes $L(\theta)$ is called the *maximum likelihood estimate*.

In many cases it is more convenient to maximize $\ell(\theta)$.

For example, suppose $x_1, x_2, \ldots, x_n$ are data observed from a $X \sim N(\mu, 1)$ population. A likelihood for μ is:

$$L(\mu) = (2\pi)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$\ell(\mu) = \log L(\mu) = C - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To maximize:

$$0 = \frac{d}{d\mu}\ell(\mu) = \sum_{i=1}^{n}(x_i - \mu) \Longrightarrow \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

the maximum likelihood estimator

A final technicality. When you replace the data $x_1, x_2, \ldots, x_n$ with its “model”, the sample: $X_1, X_2, \ldots, X_n$, inside the maximum likelihood estimate, you end up with the *maximum likelihood estimator*, or MLE.

Traditionally the MLE notation is the parameter-with-a-hat.

For example, the maximum likelihood estimator for μ using a sample from a $N(\mu, 1)$ population is:

$$\hat{\mu} = \overline{X}$$

Everything so far extends to vector parameters. For example (textbook example 9.21), the maximum likelihood estimates given data $x_1, \ldots, x_n$ from a $N(\mu, \sigma)$ population, the MLE for $\theta = (\mu, \sigma^2)$ are:

$$\hat{\mu} = \overline{X} \qquad \widehat{\sigma^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n}$$

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a 1-1 function.
4. it is asymptotically normal.
5. if $c\hat{\theta}$ is unbiased for some constant c, then $c\hat{\theta}$ is the unbiased estimator with the smallest variance (our "gold standard".)

maximum likelihood summary

The joint pmf/pdf is treated as a function of the parameter(s) θ, given the data.

This function is called a “likelihood” $L(\theta)$.

A likelihood can be thought of as the “probability” of the data.

The parameter value $\hat{\theta}$ that maximizes $L(\theta)$ is the maximum likelihood estimator.

The examples we’ve done so far have all had a closed form solution, but this isn’t necessary or even “better” in any sense.

properties of the maximum likelihood estimator

In most cases, the MLE $\hat{\theta}$ has all the following (amazing!) properties:

1. it is asymptotically unbiased.
2. it is consistent.
3. it is "invariant", which means $\widehat{h(\theta)} = h(\hat{\theta})$ when h is a smooth 1-1 function.
4. it is asymptotically normal. (Note: convergence can be slow.)
5. if $c\hat{\theta}$ is unbiased for some constant c, then $c\hat{\theta}$ is the unbiased estimator with the smallest variance, or "MVUE" (our "gold standard".)

the normal case

Population $N(\mu, \sigma)$. Observe: $x_1, \ldots, x_n$. The MLEs are (example 9.21):

$$\widehat{\theta} = \left(\hat{\mu}, \widehat{\sigma^2}\right) = \left(\overline{X}, \frac{\sum \left(X_i - \overline{X}\right)^2}{n}\right)$$

Therefore, $\overline{X}$ and S^2 are the MVUE estimators for μ and σ^2

exponential distributions - I

Population Exp(λ). Observe: $x_1, \ldots, x_n$. Let's find the MLE for $\beta = 1/\lambda$, which is the mean of the distribution.

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum x_i/\beta} \\ \ell(\beta) &= -n \log \beta - \sum x_i/\beta \\ \frac{d}{d\beta}\ell(\beta) &= -\frac{n}{\beta} + \frac{\sum x_i}{\beta^2} \end{aligned}$$

(Technicality: so the ML estimat***e*** is $\overline{x}$...)

...and the ML estimat***or*** is $\hat{\beta} = \overline{X}$. Since $E\left(\overline{X}\right) = \beta$, it is the MVUE for β.

exponential distributions - II

Recall last week when we considered estimating λ directly. We now know immediately that $\hat{\lambda} = n/(\sum X_i)$ (invariance of MLE).

Then I did all that work on the board to show:

$$E\left(\hat{\lambda}\right) = \frac{n}{n-1}\lambda$$

and that an unbiased estimator for λ was therefore

$$\frac{n-1}{n}\hat{\lambda} = \frac{n-1}{\sum X_i}$$

Now we know immediately that this is the MVUE for λ

exponential distributions - III (mind-blowing version)

I said we observed: $x_1, x_2, \ldots, x_n$. These often might be times-to-events, such as failure times of equipment, or the death/remission times of people in a medical study.

In real life analyses most stuff doesn't fail, and most people survive. Or at least we don't wait around long enought to see everything actually fail.

What we would more typically see is data as follows. "Today" I extract the historical data on the equipment I am interested in:

ID	Age	Status
A023	6.8	Failed
A324	7.2	Operating
A620	10.1	Taken Out of Service
A092	2.4	Operating
A526	5.5	Operating
A985	8.1	Failed
A723	1.5	Operating
⋮	⋮	⋮

exponential distributions - III

The model for failure times is $X \sim \text{Exp}(\lambda)$.

What is the likelihood of the data?

The likelihood for a unit to fail at time x_i is: $\lambda e^{-\lambda x_i}$

The likelihood for a unit to not have failed yet at time x_i is: $P(X > x_i) = e^{-\lambda x_i}$

For example:

ID	Age	Status	Likelihood
A023	6.8	Failed	$\lambda e^{-6.8\lambda}$
A324	7.2	Operating	$e^{-7.2\lambda}$
A620	10.1	Taken Out of Service	$e^{-10.1\lambda}$
A092	2.4	Operating	$e^{-2.4\lambda}$
A526	5.5	Operating	$e^{-5.5\lambda}$
A985	8.1	Failed	$\lambda e^{-8.1\lambda}$
A723	1.5	Operating	$e^{-1.5\lambda}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

exponential distributions - III

When the failure time is unknown, because it hasn't happened yet, we say the failure time is *censored*.

Define the *censoring indicator* c_i to be 1 if the unit failed and 0 otherwise.

Putting it all together, given times $x_1, \dots, x_n$ and censoring indicators $c_1, \dots, c_n$, the likelihood of the data is:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)^{c_i} \left(e^{-\lambda x_i}\right)^{1-c_i} \\ &= \lambda^{\sum c_i} e^{-\lambda \sum x_i} \\ \ell(\lambda) &= \log \lambda \sum c_i - \lambda \sum x_i \\ \frac{d}{d\lambda}\ell(\lambda) &= \frac{\sum c_i}{\lambda} - \sum x_i \end{aligned}$$

So $\hat{\lambda} = \frac{\sum c_i}{\sum x_i} = \frac{\#\text{ of failures}}{\text{Total Time}}$. This is called an "occurence-exposure rate".

occurrence-exposure example

Here are 50 simulated “ages” from an Exp(0.1) population, “censored” at 9.0 “years”

```
##  [1] 3.68 9.00 9.00 9.00 0.36 9.00 9.00 3.26 9.00 2.27 8.43
## [12] 9.00 9.00 9.00 9.00 9.00 6.18 9.00 9.00 0.19 2.40 9.00
## [23] 1.85 4.34 4.75 9.00 9.00 9.00 9.00 9.00 9.00 7.33 6.14
## [34] 5.07 1.35 9.00 9.00 9.00 0.04 1.79 3.81 5.97 9.00 2.50
## [45] 9.00 5.19 5.69 0.87 4.06 2.06
```

The “naive” mean life estimate (the average of the failed units only): 3.583.

The MLE: 12.583.

MLE result I published in 2016

The basic “shock and damage model” works like this:

- a unit suffers shock events that occur according to a Poisson process $N(t)$
- at each shock event, the damage suffered is X_i (in general, random, but not necessarily)
- the cumulative damage is a sum of a random number of random damages:

$$Z(t) = \sum_{i=1}^{N(t)} X_i$$

- the unit fails the moment $Z(t)$ reaches some threshold

MLE result I published in 2016

One day I encountered a situation where the company only knew the age of an item, if an item had ever suffered at least one shock event (some items never did), and the total amount of damage.

The company needed an estimate of the Poisson rate λ at which shocks occurred (among other things).

So I went looking for the method that everyone used to estimate the rate in these situations. But nobody had ever done this before.

(Many OR professors like to propose models, but often do not dirty themselves with actual data.)

MLE result I published in 2016

I introduced a “shock indicator” d_i which is 1 when one or more shocks occurred, and 0 otherwise.

The probabilities of having endured 0, or 1+ shocks by age t_i are:

$$P(N(t_i) = 0) = e^{-\lambda t_i}$$
$$P(N(t_i) > 0) = 1 - e^{-\lambda t_i}$$

The likelihood for λ is therefore:

$$L(\lambda) = \prod_{i=1}^{n} \left(e^{-\lambda t_i}\right)^{1-d_i} \left(1 - e^{-\lambda t_i}\right)^{d_i}$$
$$\ell(\lambda) = -\lambda \sum_{i=1}^{n} t_i(1 - d_i) + \sum_{i=1}^{n} d_i \log\left(1 - e^{-\lambda t_i}\right)$$

This can only be maximized numerically.

hypothesis testing

context: a one-sample t interval example

Let's suppose it is widely known ("everyone knows") that the healthy amount of `Fe` to have in engine oil is, say, "4ppm".

You have $n = 30$ haul trucks and you want to assess the health of the fleet of engines.

You take the oil samples and you end up with a sample average of $\overline{x} = 7.15$ ppm and a sample standard deviation of $s = 8.82$ ppm.

A 95% confidence interval for the mean amount of `Fe` in your fleet of engines is easily computed to be:

$$7.15 \pm t_{29,0.025}\frac{8.82}{\sqrt{30}}$$

or [3.86, 10.44].

Are your engines healhty, or not?

"Hypothesis testing" generally involves using data to make a principled statement about a parameter value.

hypotheses

Specific statements about parameter values can have practical meanings.

In the `Fe` example, the model used for the population was $X \sim N(\mu, \sigma)$.

Some statements include:

$$\mu = 4 \qquad \mu \neq 4 \qquad \sigma = 1 \qquad \sigma^2 > 5$$

To be honest, when only considering one population, this all can look mysterious and arbitrary.

A statement about a parameter value is called a *hypothesis*.

some more natural hypotheses

Consider two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$. The most obviously interesting hypothesis is:

$$\mu_1 = \mu_2$$

which is the hypothesis that encapsulates “no difference”.

Similarly, consider two population Bernoulli(p_1) and Bernoulli(p_2). We might also have:

$$p_1 = p_2$$

to mean “no difference”

null hypothesis and alternative hypothesis

In hypothesis testing we settle on two “hypotheses” concerning parameter values.

The only mathematical requirement is that they don’t contain any parameter values in common.

But in practice it is not so arbitrary.

The null hypothesis, denoted by H_0, is almost always the “no effect”/“no difference”/“status quo” parameter value.

For example, $\mu_1 = \mu_2$ and $p_1 = p_2$.

The alternative hypothesis, denoted by H_1, is usually the complement of H_0.

For example, $\mu_1 \neq \mu_2$ and $p_1 \neq p_2$.

opinion - The Myth of the “One-Sided” Alternative

Textbooks go on about “choosing” the “appropriate” alternative hypothesis, based on little more than the hopes and dreams of the experimenter.

Students are sent on wild good chases trying to guess what the textbook author/instructor is “hoping” the alternative is.

In my **opinion** this is nonsense. The alternative should almost always be the completment of the null.

(Note: this is a scientific opinion, and not a mathematical opinion.)

“classical” hypothesis testing

Our first view of hypothesis testing has a clear goal, which is to use data to make a specific decision: to either *reject* H_0 or *not reject* H_0.

Sometimes *accept* is used as a synonym for *not reject*. The book uses the phrase *fail to reject*, which I’ve never seen anywhere else.

The main thing is to avoid attaching positive or negative connotations to any of these phrases.

The method in a nutshell: assume H_0, collect a sample, and see if the sample contradicts H_0.

Motivating example. . . a pre-fabricated furniture company needs its supplier to provide doors that are 700mm wide. Does the supplier meet this target?

The model for door width will be $N(\mu, \sigma)$, with $\sigma = 0.5$ magically known for now.

classical hypothesis testing—motivating example

The null and alternative hypotheses are:

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

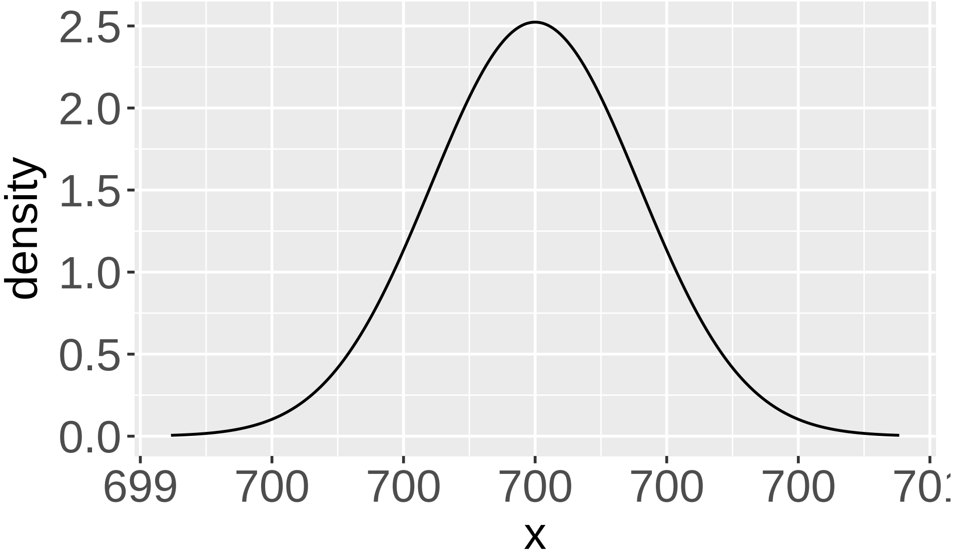
We plan to gather a sample of size $n = 10$.

What *statistic* should be used to make statements about μ. Probably a good idea to use the MLE $\overline{X}$. This is called the *test statistic*.

Suppose (temporarily, as a thought experiment) that in fact $\mu = 700$. What is the distrubtion of $\overline{X}$ and which values of $\overline{X}$ would surprise us?

$$\overline{X} \sim N\left(700, \frac{0.5}{\sqrt{10}}\right) \quad \text{The } \textit{null distribution}$$

null distribution $N(700, 0.158)$



classical hypothesis testing - the details

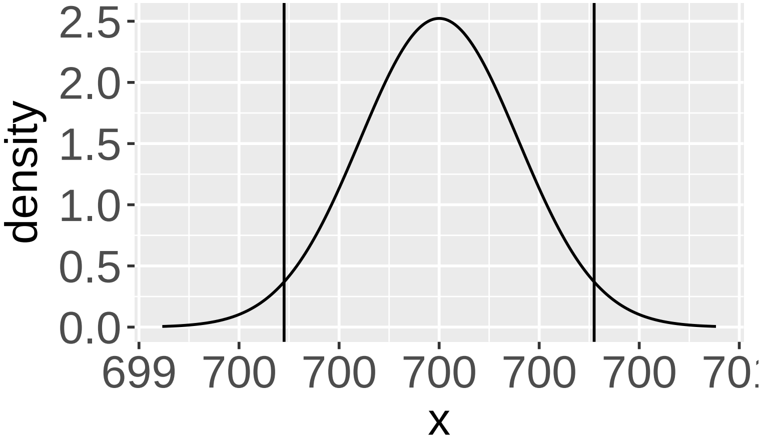
“The values that would surprise us” are defined in advance according to a pre-set probabilty α.

This is called the “size” of the test, or the “level of significance”. It is typically something small like: 0.05, 0.1, 0.05, 0.05, 0.01, 0.05, or 0.05.

α is the *probability of rejecting H_0 when it is in fact true.*

classical hypothesis testing - the details

Suppose $\alpha = 0.05$. The "area of surprise" in our motivating example is defined as $\overline{X} \leqslant 699.6901$ or $\overline{X} \geqslant 700.3099$, as in:



classical hypothesis testing - the details

The “area of surprise” is really called the “rejection region” or “critical region”.

We define two types of “error” in classical hypothesis testing:

	Action	
“Truth”	Reject	Not Reject
H_0 True	Type I Error	
H_0 False		Type II Error

$$\alpha = P(\text{Type I Error}) \qquad \beta = P(\text{Type II Error})$$

The probability $1 - \beta$ of rejecting H_0 when it is false is called the “power” of the test.

example of critical region

In our motivating example, the critical region comes from this expression that uses the null distribution:

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - 700}{0.5/\sqrt{10}} < z_{\alpha/2}\right) = 1 - \alpha$$

The region is:

$$\left\{\overline{X} < 700 - z_{\alpha/2}\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + z_{\alpha/2}\frac{0.5}{\sqrt{10}}\right\}$$

If we set $\alpha = 0.01$, say, this becomes:

$$\left\{\overline{X} < 699.593\right\} \cup \left\{\overline{X} > 700.407\right\}$$

example power calculation

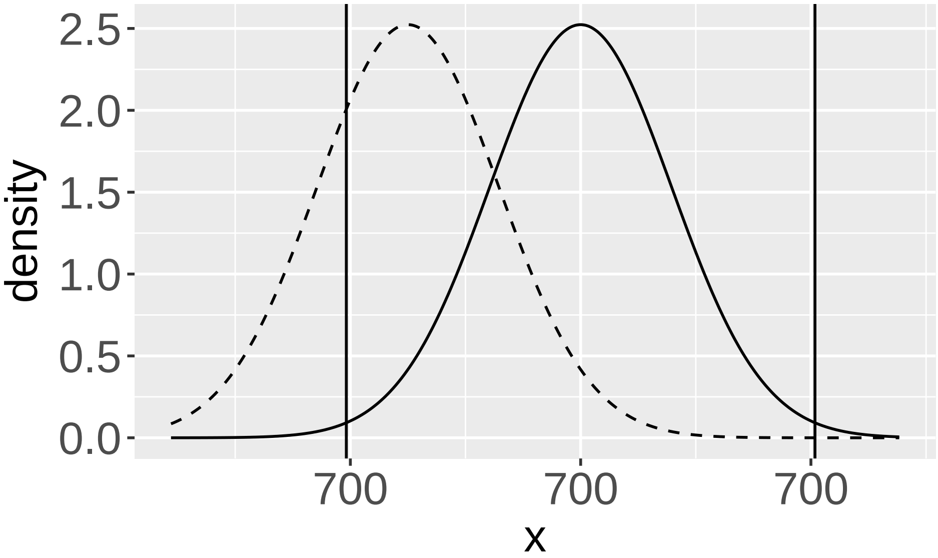
For an explicit power calculation, one needs a specific alternative.

So, suppose in fact the supplier makes doors that are $\mu_1 = 699.7$mm wide. So in fact $\overline{X} \sim N(699.7, 0.5/\sqrt(10))$

What is the probability of “rejecting H_0”?

$$\begin{aligned} P_{\mu_1}(\overline{X} < 699.593) + P_{\mu_1}(\overline{X} > 700.407) &= P(Z < -0.677) + P(Z > 4.471) \\ &= 0.249 + 0 \end{aligned}$$

power in pictures



size, power, and sample size

When the population is $N(\mu, \sigma)$ and the sample is $X_1, \ldots, X_n$ and the hypotheses are $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, the generic rejection region is, for fixed α:

$$\left\{ \overline{X} < \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} \cup \left\{ \overline{X} > \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$$

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.
4. Be a researcher who only publishes results in which H_0 is rejected.
5. Believe in the sanctity of $\alpha = 0.05$.
6. Perform as many hypothesis tests as you like on the same dataset.
7. Use "one-sided alternatives" because you think you really know what you are doing.
8. Think that you really know what you are doing.

p-values

a story that never really happened

Are the doors the right width or not?

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

Use population model $N(\mu, \sigma = 0.5)$. Set $\alpha = 0.05$. Plan to collect a sample of size $n = 10$. The rejection region is:

$$\left\{\overline{X} < 700 - 1.96\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + 1.96\frac{0.5}{\sqrt{10}}\right\}$$
$$\left\{\overline{X} < 699.69\right\} \cup \left\{\overline{X} > 700.31\right\}$$

You actually measure 10 doors. The sample average is $\overline{x} = 699.68$. We have a 2319! PUSH THE RED BUTTON!!! **REJECT THE NULL! REJECT THE NULL!**

a story that never really happened

So you cancel the contract with the supplier, who goes out of business.

It turns out that the summer student who compiled the data made a small error in recording one of the door widths - putting 700.2 instead of 700.4 for that one record.

So actually $\overline{x}$ is 699.70.

Everthing has ***completely changed***. H_0 is not rejected. FAIL TO REJECT! FAIL TO REJECT! Situation is niner-niner-zero.

But it's too late. The market has decided. Lives are destroyed. Demagogues rise to power in the wake of mass disillusionment.

Another option is to use something called a *p-value*.

p-value

A p-value is the probability (calculated using the H_0 parameter value) of observing a value of the test statistic “more extreme” that what was actually observed.

“More extreme” just means further away (in absolute value) than the H_0 parameter value.

In the Doors example, $\overline{x}$ was thought at first to be 699.68, which is 0.32mm away from 700. The probability of being *more than* 0.32mm away from 700 is:

$$P(\overline{X} < 699.68) + P(\overline{X} > 700.32) = 0.021 + 0.021 = 0.043$$

After the correction, $\overline{x}$ is now 699.70. The p-value is now:

$$P(\overline{X} < 699.70) + P(\overline{X} > 700.30) = 0.029 + 0.029 = 0.058$$

Too bad you didn’t know about p-values before causing WWIII.

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is “small enough.” So don’t ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be “no evidence”
- 0.063 might be “weak evidence”
- 0.0031 might be “evidence”
- 0.0000014 might be “strong evidence”
- 3×10^{-15} might be “overwhelming evidence”

Think in terms of orders of magnitude.

things that annoy me

When people ask me how small a p-value “has” to be.

When people compute a p-value, and then say “The p-value is smaller than 0.05, so I reject the null hypothesis.”

$100 \cdot (1-\alpha)\%$ C.I. versus hypothesis test with size α

A C.I. formula and a rejection region formula for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ will be based on (something like):

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

For the C.I., unwrap to isolate μ in the middle. For the R.R., put $\mu = \mu_0$ and unwrap to isolate $\overline{X}$ in the middle.

The following is true:

$$\text{Reject } H_0 \text{ at level } \alpha \qquad \Longleftrightarrow \qquad 100 \cdot (1-\alpha)\% \text{ C.I. does not contain} \mu_0$$

the “one-sample t test”

You’ll never know σ, so use the data to estimate σ with s, as usual.

From the Doors example where $n = 10$ and $\overline{x} = 699.70$, suppose also that that $s = 0.389$.

The p-value is now calculated based on:

$$\frac{\overline{X} - 700}{s/\sqrt{n}} \sim t_9$$

$$P(\overline{X} < 699.70) + P(\overline{X} > 700.30) = P(t_9 < -2.441) + P(t_9 > 2.441) = 0.019 + 0.019 = 0.037$$

the “two-sample t-test”

A more realistic hypothesis testing scenario.

Two populations: $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. The obvious hypotheses will always be:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

The “parameter” is $\theta = \mu_1 - \mu_2$, estimated (as usual) by $\overline{X_1} - \overline{X_2}$ from samples of sizes n_1 and n_2.

Two possibilities:

$$\frac{\overline{X_1} - \overline{X_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad \text{or} \qquad \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

two-sample t-test example

Modified from 10.106. Can nutritional counselling change blood cholesterol level? A group of 15 people received counseling for 8 weeks. A group of 18 people did not.

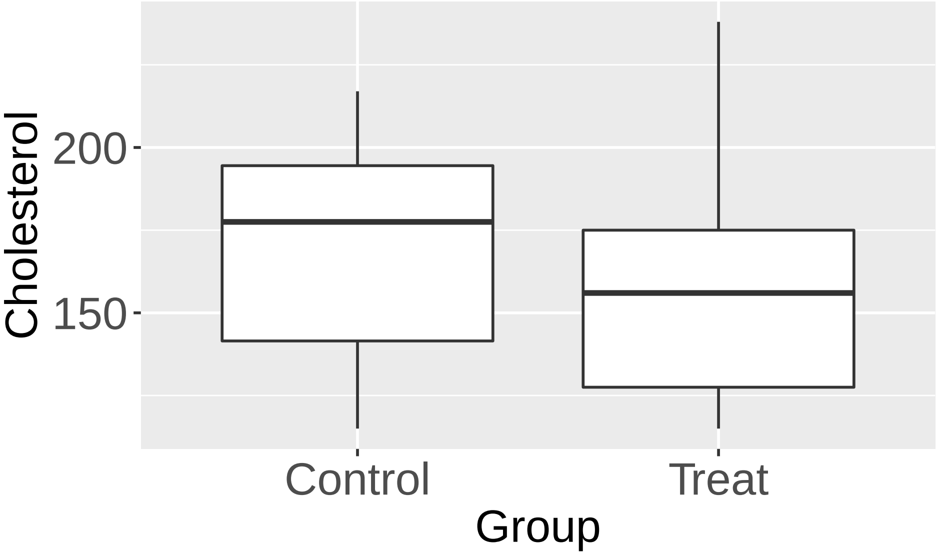
The readings are made available by the textbook in the following terrible manner:

two-sample t-test example

A Real Dataset:

```
##      Group Cholesterol
## 1    Treat         129
## 2    Treat         131
## 3    Treat         154
## 4    Treat         172
## 5    Treat         115
## 6    Treat         126
## 7    Treat         175
## 8    Treat         191
## 9    Treat         122
## 10   Treat         238
## 11   Treat         159
## 12   Treat         156
## 13   Treat         176
## 14   Treat         175
```

two-sample t-test example - plot



two-sample t-test example - equal variance version

```
## # A tibble: 2 × 4
##     Group     n  X_bar      S
##    <fctr> <int>  <dbl>  <dbl>
## 1 Control    18 170.00 30.788
## 2   Treat    15 156.33 33.090
```

```
## 
##  Two Sample t-test
## 
## data:  Cholesterol by Group
## t = 1.23, df = 31, p-value = 0.23
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.0417 36.3750
## sample estimates:
## mean in group Control   mean in group Treat
##                170.00                156.33
```

two-sample t-test example - no variance assumption version

```
## 
##  Welch Two Sample t-test
## 
## data:  Cholesterol by Group
## t = 1.22, df = 29, p-value = 0.23
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.2584 36.5917
## sample estimates:
## mean in group Control   mean in group Treat 
##                170.00                156.33 
```

the “two-sample t-test”

A more realistic hypothesis testing scenario.

Two populations: $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. The obvious hypotheses will always be:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

The “parameter” is $\theta = \mu_1 - \mu_2$, estimated (as usual) by $\overline{X_1} - \overline{X_2}$ from samples of sizes n_1 and n_2.

Two possibilities:

$$\frac{\overline{X_1} - \overline{X_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad \text{or} \qquad \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

two-sample t-test example

Modified from 10.106. Can nutritional counselling change blood cholesterol level? A group of 15 people received counseling for 8 weeks. A group of 18 people did not.

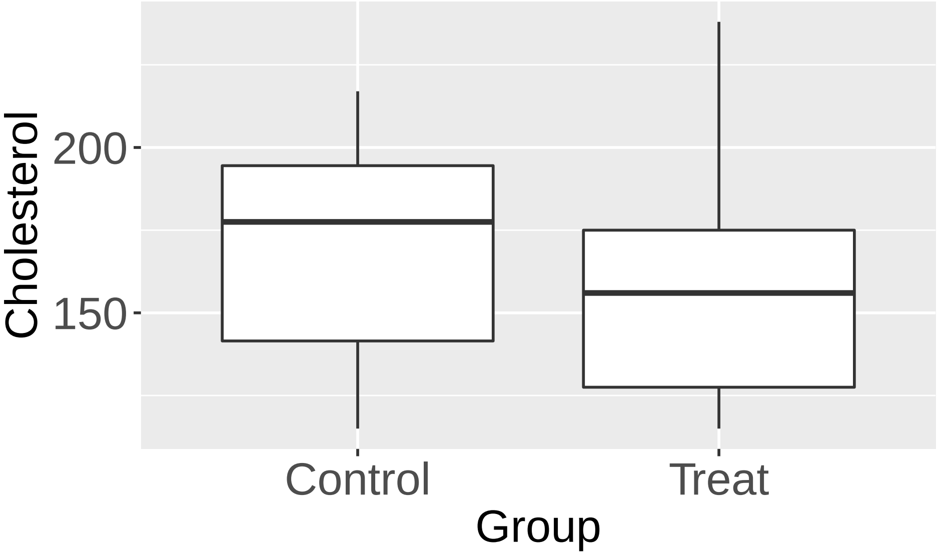
The readings are made available by the textbook in the following terrible manner:

two-sample t-test example

A Real Dataset:

ID	Group	Cholesterol
22	Control	195
24	Control	198
20	Control	132
11	Treat	159
23	Control	188
32	Control	217
14	Treat	175
26	Control	168
15	Treat	126
35	Control	140
30	Control	208
21	Control	196
25	Control	187
1	Treat	129
5	Treat	115
27	Control	115
2	Treat	131
8	Treat	191
34	Control	193
4	Treat	172
3	Treat	154
12	Treat	156
28	Control	165
36	Control	146
7	Treat	175
31	Control	133
13	Treat	176
19	Control	151
10	Treat	238

two-sample t-test example - plot



two-sample t-test example - equal variance version

Group	n	X_bar	S
Control	18	170.00	30.79
Treat	15	156.33	33.09

```
##
##  Two Sample t-test
##
## data:  Cholesterol by Group
## t = 1.23, df = 31, p-value = 0.23
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.0417 36.3750
## sample estimates:
## mean in group Control   mean in group Treat
##                170.00                156.33
```

two-sample t-test example - no variance assumption version

Group	n	X_bar	S
Control	18	170.00	30.79
Treat	15	156.33	33.09

```
##
##  Welch Two Sample t-test
##
## data:  Cholesterol by Group
## t = 1.219, df = 29.04, p-value = 0.233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.25835 36.59169
## sample estimates:
## mean in group Control   mean in group Treat
##               170.000               156.333
```

two-sample t-test example - or is it?

Question 10.54. Nine people had breathing rates measured with and without elevated CO levels.

Subject	WithCO	WithoutCO
1	30	30
2	45	40
3	26	25
4	25	23
5	34	30
6	51	49
7	46	41
8	32	35
9	30	28

Does CO impact breathing frequency?

two-sample t-test example - or is it?

Two populations are $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

The two samples are $X_{11}, \ldots, X_{19}$ and $X_{21}, \ldots, X_{29}$.

But they are surely not independent. We should examine the differences $D_1, \ldots, D_9$, which will be $N(\mu_D, \sigma_D)$ where $\mu_D = \mu_1 - \mu_2$.

Here's a one-sample case where the null and alternatives are actually self-evident:

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

two-sample t-test example - or is it?

The analysis:

n	X_Bar_1	X_Bar_2	S_1	S_2	X_bar_D	S_D
9	35.444	33.444	9.462	8.502	2	2.55

```
##
##  One Sample t-test
##
## data:  co$WithCO - co$WithoutCO
## t = 2.353, df = 8, p-value = 0.0464
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.0402733 3.9597267
## sample estimates:
## mean of x
##         2
```

single proportion example (something funny happens)

From the second test, that gas company "knew" the proportion of defective meters was 0.01. Let's change that to "assumes" (perhaps based on some industry knowledge). As usual, the single sample scenarios tend to be a bit contrived.

Work Team Beta inspects 2000 meters and finds 24 defective ones. Is there evidence that the company's assumption is inaccurate?

$$H_0 : p = 0.01$$
$$H_1 : p \neq 0.01$$

We'll use the MLE $\hat{p}$, for which we know:

$$\hat{p} \sim^{approx} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

To calculate the p-value (or to get a critical region) we plug the H_0 value to obtain the null distribution. This happens to eliminate the unknown variance problem!

single proportion example

We observe $\hat{p}_{obs} = 0.012$. What is the p-value?

$$\begin{aligned} P(\hat{p} < 0.008) + P(\hat{p} > 0.012) &= P\left(Z < \frac{0.008 - 0.01}{\sqrt{\frac{0.01(1-0.01)}{2000}}}\right) + P\left(Z > \frac{0.012 - 0.01}{\sqrt{\frac{0.01(1-0.01)}{2000}}}\right) \\ &= P(Z < -0.899) + P(Z > 0.899) \\ &= 0.369 \end{aligned}$$

two proportion example - a little trick

Much more natural.

Let's say Work Team Beta found $x_1 = 24$ defective meters in $n_1 = 2000$ inspections, and Work Team Delta found $x_2 = 14$ in $n_2 = 1500$ inspections. Do the teams find defectives at the same rate?

We are comparing a Bernoulli(p_1) with a Bernoulli(p_2). The null and alternative are self-evident:

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

We will use $\hat{p}_1 - \hat{p}_2$, which we know satisfies:

$$\hat{p}_1 - \hat{p}_2 \sim^{approx} N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

two proportion example - null distribution little trick

When computing the p-value, we plug in the H_0 fact that $p_1 = p_2$, which we will denote by just p. The variance of the "null distribution" reduces to:

$$p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

We don't know p. Use the data, which, under the null hypothesis, are just 0's and 1's from the same Bernoulli(p) distribution. We pool them together to get:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

So the null distribtion is:

$$\hat{p}_1 - \hat{p}_2 \sim^{approx} N\left(0, \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$$

two proportion example

In our example we had $x_1 = 24$, $n_1 = 2000$, $x_2 = 14$, $n_2 = 1500$. So:

$$\hat{p} = 0.010857$$

and the standard deviation of the null distribution is 0.00354.

Also, $\hat{p}_1 - \hat{p}_2 = 0.002667$

The p-value is 0.451228 based on $2P(Z < -0.75337)$