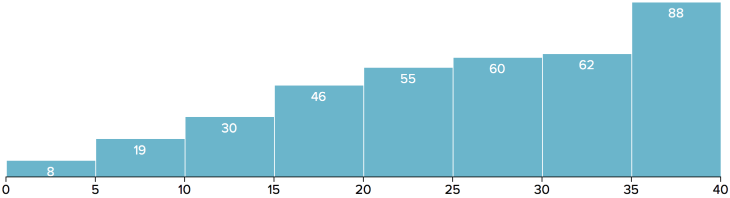
Probability

Kenneth A. Ribet

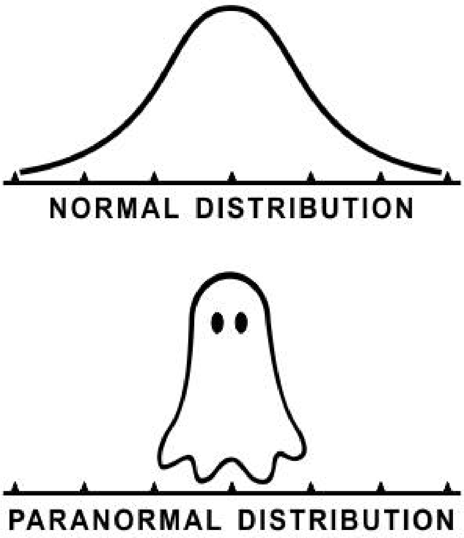
Math 10A
October 32, 2016

Midterm results

- 368 students took the exam.
- Scores range from 0 (six students) to 40 (16 students). The histogram of scores looks like a flight of stairs:



- Mean was 25.61.
- Median was 27.0.
- Standard Deviation $= 9.9$.


NORMAL DISTRIBUTION
PARANORMAL DISTRIBUTION

Meals

Breakfast on Thursday at 8:30AM. There may be a slot or two remaining. (There are currently 18 people signed up, not including me.)

Pop-in lunch on Friday, November 4 at 12:30PM.

Pop-in lunch on Monday, November 7 at 12:10PM.

Why these slides?

They provide a written record of the class meeting so that we can look back and see what we discussed.

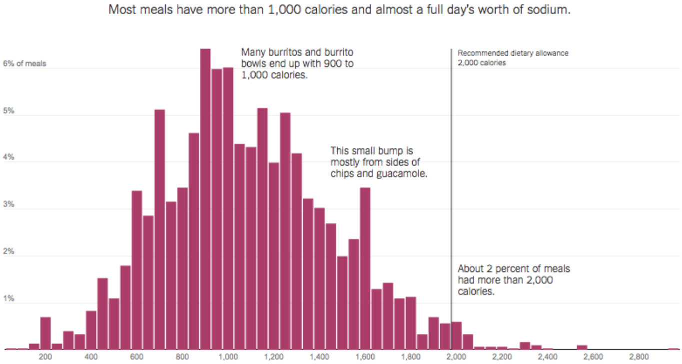
Typing the slides helps me prepare the lecture.

They *might* get projected during the class.

We can switch to the chalkboards at any time!

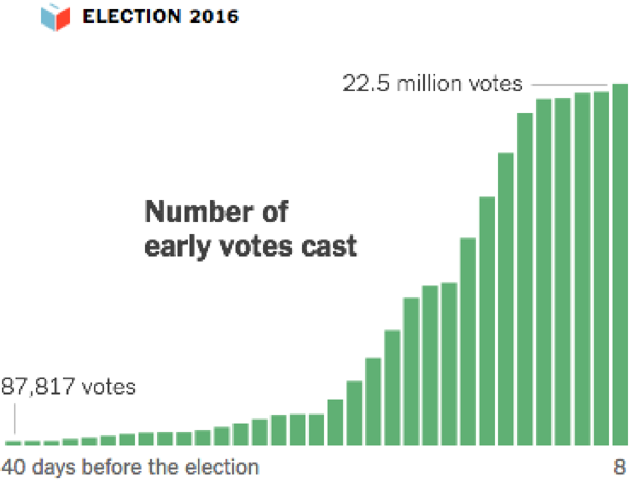
Histograms

Histograms are pictures like the “flight of stairs” image that represents the grade distribution. They are used for pictorial representations of data.



Here’s another historgram (“At Chipotle, How Many Calories Do People Really Eat?,” NY Times).

Now here's a picture from the NY Times that looks superficially like a histogram. In fact, it reminds me a bit of the MT2 histogram:



In fact, this last image displays the *total* number of early votes as it varies over time. The information is *cumulative*. The bars are rising—by definition. I think that many people use the phrase “cumulative histogram” for images like this.

If you had a histogram depicting the number of early votes cast each day, you could construct the NY Times’s cumulative historgram by summation: the height of the green bar for a given day would be the sum of the heights of the bars of the histogram for that day and all previous days.

PDFs, CDFs

PDFs are probability density functions, *not* `.pdf`'s (documents).

CDFs are cumulative density functions.

The idea is that we model these bar pictures ("discrete") wth continuous functions. The function that models a historgram is basically a PDF and the function that models the cumulative display is then the integral of the PDF. This integral is a CDF.

Technically, a PDF is a function $f(x)$ (x a real number) such that:

- $f(x) \geq 0$ for all x;
- the total area under $f(x)$ is 1: $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

The second condition is a normalization condition. We're not trying the model the number of students who got grades between 25 and 30, but rather the fraction of the class getting grades in that range. The sum of all the fractions is 1.

Usually we assume that $f(x)$ is piecewise continuous, meaning that it is basically continuous except that it's allowed to have a few breaks here and there.

Suppose that $g(x)$ is positive and that the total area under $g(x)$ is a non-zero number A. Then $\frac{1}{A}g(x)$ is a PDF.

Example: if $g(x) = e^{-x^2}$, then $A = \sqrt{\pi}$, as we saw together last week. Therefore $\frac{1}{\sqrt{\pi}}e^{-x^2}$ is a PDF.

The idea is the PDFs can be used to model a data set. Imagine that we have 1,000,000 scores (for something) that are distributed between $-\infty$ and $+\infty$. If the fraction of scores between a and b is roughly $\int_a^b f(x)\,dx$, then the PDF $f(x)$ is a good fit for the data set.

Probability from a mathematical point of view

The basic notion is that of a *probability space*. This is a set Ω and a function P on the set of subsets of Ω such that:

- $P(A) \in [0, 1]$ for all $A \subseteq \Omega$;
- $P(\Omega) = 1$, $P(\emptyset) = 0$;
- $P(A \cup B) = P(A) + P(B)$ is A and B are disjoint.

It is a white lie that P is defined on *all* subsets of Ω. For example, if $\Omega = \mathbf{R}$, then $P(A)$ is not defined for A sufficiently quirky.

See also the discussion beginning on page 12 of `Prob-Stat.pdf`.

A typical Math 10B example is to take Ω to be finite and to set $P(A) = \frac{|A|}{|\Omega|}$. Then every point of Ω gets the same probability, namely $\frac{1}{|\Omega|}$. The probability is then said to be *uniform*.

In Math 10A, we are interested in infinite spaces like **R** or the unit interval $[0, 1]$. In that situation, we can't count! We say that we are in the *continuous* case.

We can introduce probability for the real line by choosing a PDF $f(x)$. Recall that this means that $f(x)$ is non-negative and that the total area under $y = f(x)$ is 1. If A is a set of real numbers, we define: $P(A) = \int_A f(x)\,dx$.

For example, if $A = [a, b]$, then $P(A) = \int_a^b f(x)\,dx$.

One might say things like: “The probability that x lies between -7 and 4 is
$$\int_{-7}^{4} f(x)\,dx.\text{”}$$

More generally, if A is a set of numbers, we introduce this function $g(x)$: $g(x) = 1$ for $x \in A$ and $g(x) = 0$ for $x \notin A$. Then, by definition,

$$\int_A f(x)\,dx = \int_{-\infty}^{\infty} f(x)g(x)\,dx.$$

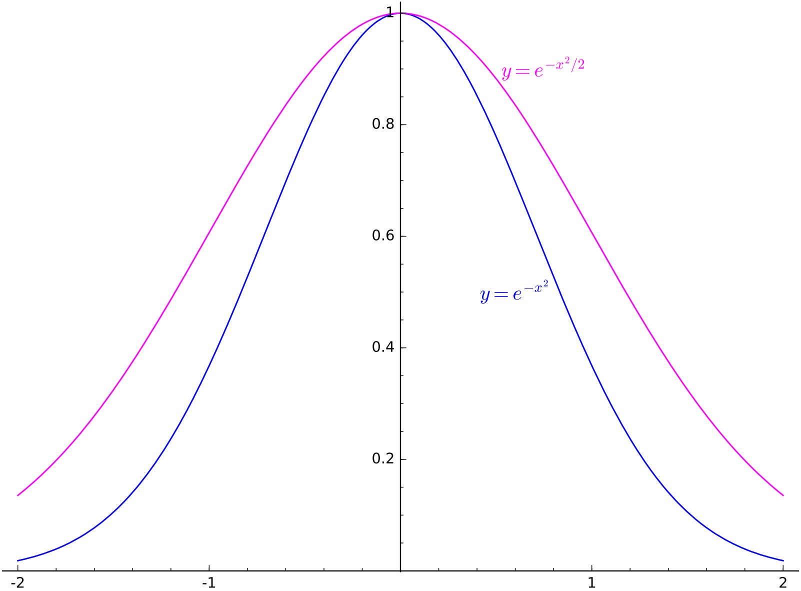
Some minutes ago, we considered this PDF: $f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$.

Here's a mild generalization:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The term μ serves to slide the function to the right by μ when μ is positive or to the left by $-\mu$ when μ is negative. The parameter σ compresses or uncompresses the graph of the function.

In the initial example, $\sigma^2 = \frac{1}{2}$, $\sigma = \frac{1}{\sqrt{2}}$.



1
0.8
0.6
0.4
0.2
-2
-1
1
2
$y = e^{-x^2/2}$
$y = e^{-x^2}$

CDFs

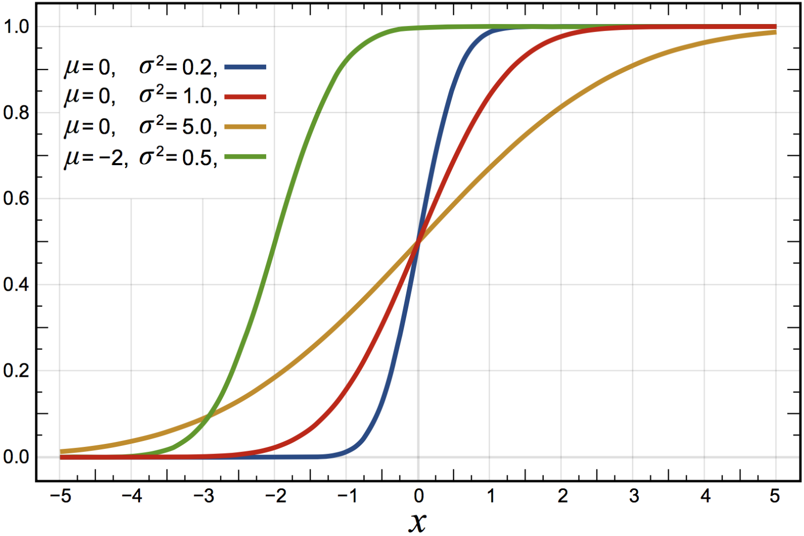
The cumulative distribution function associated with the PDF is the function $F(x)$ defined as follows: $F(x)$ is the probability that a number is $\leq x$; i.e., $F(x) = P((-\infty, x])$. In symbols:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

Note that $F(x)$ is an antiderivative of $f(x)$: by the fundamental theorem of calculus,

$$F'(x) = f(x).$$

Thus, for example, $P([a, b]) = \displaystyle\int_a^b f(x)\,dx = F(b) - F(a)$ (which we could have figured out directly because the probability of being in $[a, b]$ is the difference of the probabilities of being in $(-\infty, b]$ and being in $(-\infty, a]$).



CDFs for some Gaussian PDFs. See Wikipedia for more info.

If Ω is a probability space, a function

$$X : \Omega \to \mathbf{R}$$

is called a *ramdom variable*.

A random variable is not random.

A random variable is not a variable.

A random variable is a function.

The point of random variables is that probability spaces Ω often have little to do with numbers. The points of Ω might be sequences of data. They might be frogs. They might (as it happens) be numbers. It's often helpful to think of them as possible outcomes of some experiments.

Let X be a random variable For $B \subseteq \mathbf{R}$, we write "$X \in B$" for the event

$$\{\, x \in \Omega \,|\, X(\omega) \in B \,\}.$$

The number $P(X \in B)$ is then described as "the probability that the value of X is in B."

Consider this example: you flip a coin 10 times and think of the outcome of this experiment as a string of length 10 where each character is `H` or `T`. The probability space Ω is the set of all 2^{10} possible outcomes. For example, one of the outcomes is `HHHTTTHTHT`. We'd normally consider all outcomes to be equally likely (unbiased coin). Each string is then assigned the same probability $1/2^{10}$ and an event $A \subseteq \Omega$ gets the probability $|A|/2^{10}$.

One asks questions like “What’s the probability that we get more than two heads?" For this question, we’re getting back to numbers. Technically, we have the function

$$X : \Omega \to \mathbf{R}$$

that sends each string to the number of $\mathtt{H}$s in the string. The probability of the event

$$A = \{\, \omega \in \Omega \,|\, X(\omega) > 2 \,\}$$

is the probability of getting more than two heads. This probability is $\dfrac{|A|}{2^{10}}$, so answering the question amounts to counting the number of elements in A.

You’ll learn how to do that in 10B, by the way.