AGAINST ALL ODDS
EPISODE 3 – "HISTOGRAMS"
TRANSCRIPT

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

INTRO

Pardis Sabeti
Hello, I'm Pardis Sabeti and this is *Against All Odds*, where we make statistics count.

Like most people, I'm scared of getting hit by lightning. And while it's against the odds, it isn't against all odds – hundreds of people are struck by lightning every year in the United States, and dozens are killed. What's more, fires started by lightning strikes cause hundreds of millions of dollars of property damage. The old wives' tale is wrong – lightning can definitely strike twice, and sometimes even a whole lot more often than that!
Take a look at this map of Colorado, for instance, which shows over seven million lightning strikes in the years from 1989 to 2005 – that's about half a million bolts of lightning a year.

Meteorologist Raul Lopez began collecting detailed data on lightning strikes in Colorado back in the 1980s.

> **Raul Lopez**
> When we first started to get the data we were overwhelmed by the amount of these data. In the past, researchers had been looking at just a handful of lightning flashes, maybe over several years a hundred, two hundred. Here, we collected, in one year we collected three quarters of a million flashes in just a limited area here in Colorado. So we were overwhelmed and I said, well, what do we do with it? So the first thought was, well, where does it occur, when does it occur, and how much?

Pardis Sabeti
When you are overwhelmed by data, it's statistics to the rescue, because by figuring out a way to display your data you can see patterns and make inferences about what those data mean – patterns that are otherwise simply buried and lost in all those numbers.

The statistical tool Raul Lopez turned to was a special form of bar graph called a histogram. His first goal was to find out when most lightning storms start, so he made a graph with time of day along the horizontal axis, and percent of days with first flash at the time on the vertical axis.
So in this histogram, each bar represents one hour, and its height is the percentage of days that had their first lightning flash within that hour.

Now you can immediately see two very striking features of this histogram. First, it's roughly symmetrical around the tallest bar, which tallies the flashes between 11 am and noon. Of course it's not perfectly symmetrical, but that's the way it usually is with real data. The second thing that leaps out is how tightly the time of first strike clusters around the central bar, with the range from 10am to 1pm

accounting for most of the days' first strikes. And there are no first strikes at night.

> **Raul Lopez**
> The interesting thing about this histogram showing the clustering was that, look, there is something regular here, something that tends to occur very frequently during the summer. A lot of these days the first flashes are tending to occur at a given time. Immediately you ask, well, why? Why is that? And the thing is that here in this mountainous region, the topography induces some circulation; and because of the circulation, you have areas where clouds would form more than in others, and at some times more than other times.

Pardis Sabeti
What causes those clouds to form are winds from the eastern plains carrying warm moist air. When the wind hits the mountains it’s forced upward where it meets and mixes with colder air higher in the atmosphere. And this is a regular daily occurrence during the Colorado summer.

Lopez and his colleagues next looked at the time of day when the maximum number of lightning flashes were recorded.
They found a similar pattern, with a peak showing that most flashes occur between four and five in the afternoon.

But there's one big difference from the first flash histogram. Look at this – on a few days the maximum was in the early hours of the morning. Remember, data points like these, that stand out from the overall pattern of the distribution, are called outliers. And outliers are often the most intriguing features of a histogram, practically begging for an explanation.

> **Raul Lopez**
> When you look at the distribution of the maximum activity, you notice that there’s some that were not clustering at all but were occurring far away from the center. You immediately tend to think about outliers, and why.

Pardis Sabeti
The explanation Lopez and his colleagues came up with for the outliers was that they occur on days when larger weather systems, specifically very strong winds from fast moving weather fronts, overpower the local effect.

So here's a great example of how a histogram, by presenting the data as a picture, can reveal things that otherwise might be missed, and what's more, make you think about what the data mean in new way.

Since the pioneering work of Raul Lopez and his colleagues on Colorado lightning, the data collection has continued. Here is a map showing all 7 million of

those lightning flashes from 1989 to 2005, plotted through the day. It beautifully confirms Lopez's histogram showing the peak activity in the late afternoon.

And I can't resist showing you this, a histogram produced just recently, plotting the number of people injured or killed by lightning strikes in the last thirty years. It shows the same clustering as Raul Lopez's histograms, but interestingly, the peak time for getting struck by lightning is around 2pm, about midway between the peaks of the first strike and maximum activity histograms. Explanation anyone?

Before leaving histograms, I want to make a couple more points. The first is that it's very important when plotting a histogram to choose the best class size – that is, the width of your intervals along the horizontal axis. Lopez chose one hour for his data, and it works well. But suppose I wanted to make a histogram of the density of workday traffic passing near my office on the Massachusetts Turnpike. I could choose a four-hour interval, producing a histogram like this, with 6 bars – not all that informative. Moving to a two-hour interval is better... and an hour is better still in showing the peak traffic flow.

But what if I wanted to get even finer-grained, say 5 minutes? It's pretty unwieldy. There's so many bars it's hard to focus in on individual time intervals. In some ways, the histogram is now less informative, not more. You want to strike a balance--enough bars to make sense of what's going on, and to recognize patterns; not so many bars that you're lost again in the details.

So we've seen how histograms can show at a glance the essence of a whole lot of numbers. Here's one last example, a plot of the weekly wages of U.S. workers in the year 1992. See how it's skewed, with most people earning around $450 a week? As you go out to what's called the tail of the distribution, here to the right, the salaries get bigger, but the numbers get smaller. Statisticians say a distribution like this is skewed to the right, because the right side of the histogram extends much further out than the left side.

Now here's a histogram of the same data for the year 2011. Look what's happened. The skew has become more pronounced, and the tail has grown longer. Suddenly our little discourse on histograms could become highly political!

I'm Pardis Sabeti for *Against All Odds*. See you next time!

PRODUCTION CREDITS

Host – Dr. Pardis Sabeti

Writer/Producer/Director – Maggie Villiger

Associate Producer – Katharine Duffy

Editors – Brian Truglio
- Jared Morris
- Seth Bender

Director of Photography - Dan Lyons

Audio – Timothy Wessel

Sound Mix: Richard Bock

Animation – Jason Tierney

Title Animation – Jeremy Angier

Web + Interactive Developer – Matt Denault / Azility, Inc.

Website Designer – Dana Busch

Teleprompter – Sue Willard-Kiess

Hair/Makeup – Emily Damron

Additional Footage
- Pond5/Francois Arseneault
- Pond5/Michael Vorobiev
- Pond5/MovingImages
- Pond5/Will Schmidt
- iStock/Dave Parsons
- Pond5/Ozgur Cagdas
- Pond5/Jackson Kitchell
- National Park Service

Additional Sound:
- Freesound.org/Erdie
- Freesound.org/Herbert Boland
- Freesound.org/RHumphries
- Freesound.org/Dynamicell
- Freesound.org/dobroide

- Freesound.org/hantorio
- Freesound.org/artifact
- Freesound.org/Andy_Gardner
- Freesound.org/FunnyMan374
- Freesound.org/mangiabambini
- Freesound.org/guitarguy1985
- Freesound.org/Cyril Laurier
- Freesound.org/ERH
- Freesound.org/cognito perceptu
- Freesound.org/cinemafia

Music
DeWolfe Music Library

Based on the original Annenberg/CPB series *Against All Odds*,
Executive Producer Joe Blatt

Annenberg Learner Program Officer – Michele McLeod

Project Manager – Dr. Sol Garfunkel

Chief Content Advisor – Dr. Marsha Davis

Executive Producer – Graham Chedd

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

For information about this, and other Annenberg Learner programs, call 1-800-LEARNER, and visit us at www.learner.org.