AGAINST ALL ODDS
EPISODE 9 – "CHECKING ASSUMPTION OF NORMALITY"
TRANSCRIPT

FUNDER CREDITS

Funding for this program is provided by Annenberg Learner.

INTRO

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

Examples of Normal distributions can be found in all sorts of settings. Remember, Normal curves are symmetric with that iconic single peak, what we call bell-shaped. But in the real world, data can be a lot messier than the idealized examples you might find in a textbook. It can be pretty tough simply to eyeball whether or not data are normally distributed – at least with much confidence. And down the road it's going to be very important to feel confident in our assumption that a particular distribution is in fact Normal, because that's a prerequisite for some more advanced statistical techniques. Let's get started by visiting a place where there are plenty of distributions for us to analyze… an egg farm.

Pete and Gerry's Organic Eggs is a family farm in northern New Hampshire. Most egg suppliers these days are vast factory farms. But this place is different. You won't find any battery cages at Pete & Gerry's, according to fourth generation farmer Jesse LaFlamme. The hens are well-treated and able to walk around indoors and out.

> **Jesse LaFlamme**
> So, the big thing is certified organic grain and humane treatment of the hens is what we do. And in our case we have a partnership with 35 family farms. It's providing an opportunity for smaller farms to exist and younger farmers to get into farming because agriculture is so big now, it's hard for anybody to get started in it.

Pardis Sabeti
Every morning the hens wake up in their community nests and most are ready to lay an egg. Conveyor belts collect the eggs from Pete & Gerry's 180,000 hens and gently funnel them into the onsite packaging center. There they join the eggs that get trucked in daily from the partner farms' additional 700,000 birds. All in all, we're talking about a lot of eggs since most of the hens are laying at a rate of one a day!

> **Jesse LaFlamme**
> It's completely random sizes out of the nest, they haven't been washed, they haven't been weighed, checked for cracks, or anything.

Pardis Sabeti
All that comes next. Machines wash and scrub the eggs clean, and cameras automatically check for any dirty ones that slipped through. An acoustic crack detection process shunts aside any damaged eggs. Once they've made it this far, the remaining eggs are weighed.

Jesse LaFlamme
So, every egg in this grid has an address, so this may be packer number 16, solenoid 7, and it's destined to go there through the process.

Pardis Sabeti
The computer automatically sorts each egg into a carton with others of similar size and quality.

Jesse LaFlamme
And if everything goes well, you can see, no one ever physically touches the eggs.

Pardis Sabeti
They're packed up and shipped out to markets all along the east coast.

Plenty of data get collected at Pete and Gerry's… How much water and feed the birds are eating. How they're growing. How many and what sizes of eggs they're laying. How production is running on the packaging line. Are any of the data distributions they see Normal? Let's start by taking a look at the weights of 7-week-old hens in one flock.

Jesse LaFlamme
So, we get them as day-old chicks and they're essentially all the same size, obviously all the same age…but as they grow it's our test to take a random sample from the barn—usually about a hundred, a hundred pullets as we call young hens, and take their weights and then we graph it…so we expect a bell curve, Normal distribution and we have sort of a target for the middle of that curve where the hens should be. And we just don't want much more than a 10% deviation on either side of that or we know there's a problem.

Pardis Sabeti

The histogram for this flock does appear to be normal, with that one peak in the middle here, around 550 grams. A normal curve drawn over the histogram seems to be a pretty good fit. It's important to consider the bin size when we're eyeballing a histogram to see if the data are Normal. Sometimes changing the class size along the *x*-axis can really change the way the histogram looks and what once appeared perfectly bell-shaped now looks quite different.

A second option as we assess normality is to use those same hen weights to construct a boxplot. Boxplots can act as another graphical display test to see if our data are normally distributed. You want to compare the boxplot's stretch for each quarter of the data. If a distribution is normal, you'd expect to see the box containing the middle 50% of the data pretty tightly grouped in the center of the distribution, with longer whiskers indicating the increased spread of the upper and lower quarters of the data.

Take a look at the way a truly Normal distribution translates into a boxplot and compare it with our hens. The weight distribution does appear to be approximately Normal, with the whiskers each longer than the Q_1 to median distance and the median to Q_3 distance.

Going beyond eyeballing these kinds of graphic displays, there's another, more precise, way to check whether a distribution is Normal. Statisticians use software to draw what's known as a normal quantile plot. The basic idea is that you're comparing the ordered data values you have with the values you would expect for a standard Normal distribution. If your observations are normally distributed, the quantile plot points will fall close to a straight line. Since a computer will do the work for you, it's less important to understand the steps taken to construct the Normal quantile plot than it is to know how to interpret it.

Here's the Normal quantile plot for the hen weights. Our observed weights are on the *x*-axis and the expected values are on the *y*-axis. The pattern of dots in the plot lies close to a straight line, so we can conclude that our data are in fact Normal.

There's less ambiguity here than when we're looking at the histogram of the data and trying to decide if the pattern looked bell-shaped. LaFlamme should be pleased with how these particular hens are growing.

Let's try out these tests with another size range… the sizes of all the eggs collected one day at Pete & Gerry's. The USDA's grading system is based on weight and it ranges from peewee, on the tiniest extreme, through small, medium, large and extra large all the way up to jumbo at the top. LaFlamme adds an uppermost category to the size scale that chicken farmers call "had to hurt" but that's not officially recognized! We can genteelly call these whopper eggs Super Jumbo.

In the wild, the size of one bird species' eggs is something that you'd expect to exhibit a normal distribution. It's a little challenging to decide if the Pete & Gerry's size distribution looks normal. Could this be an example of messy real-world numbers? Or has the farm's careful control over hatching and breeding had an influence on the size range?

Let's investigate using the normal quantile plot for these data. The middle looks like what you'd expect for Normal data, but the tails of the distribution don't. That lower tail shows us we have more eggs at the lower range than we would expect to see if the distribution were actually Normal. We can conclude that the size distribution of eggs from Pete and Gerry's is not Normal, at least on the day these data were collected. That's logical if you understand that the size of the eggs a chicken lays increases over her lifetime.

Jesse LaFlamme

They start laying a lot of eggs very quickly, but most of those eggs are small in size. Very few large eggs, very few extra-large or jumbo. Mostly the pee-wee eggs, small eggs, medium eggs…but very quickly, over the next four weeks—what we call the case weight—increases. And that will end up yielding mostly large eggs and extra large eggs. Probably close to 90% large and extra large by the time they're 20- to 26-weeks-old.

Pardis Sabeti

The egg business has seasonal cycles – for instance, sales increase for the year-end holidays when people are stocking up for their baking needs. Pete & Gerry's tries to prepare for those cycles, knowing the age of the hens they'll need laying to meet that heightened demand with the most desirable egg sizes – large and extra large. On the day these data were recorded, there were more younger flocks laying smaller eggs… gearing up for the peak season in a couple of months' time.

As you get more familiar with Normal quantile plots, you might start to recognize predictable patterns for various non-Normal distributions. For instance, here's a histogram showing a more extreme skew to the right with a majority of very young birds laying smaller eggs. The Normal quantile plot in this case is decidedly curved and looks concave down. On the other extreme, here's how things might look for a weight distribution of eggs laid by much older chickens. The histogram would be skewed to the left because of the predominance of heftier eggs and the normal quantile plot would look curved, this time concave up.

I guess all that's just another way to say, "Don't count your chickens before they hatch" – or really, don't count your eggs before they're laid – at least when it comes to assuming data are Normal!

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

PRODUCTION CREDITS

Host – Dr. Pardis Sabeti

Writer/Producer/Director – Maggie Villiger

Associate Producer – Katharine Duffy

Editor
–Seth Bender

Director of Photography - Dan Lyons

Sound Mix – Richard Bock

Audio – Timothy Wessel

Animation – Jason Tierney

Title Animation – Jeremy Angier

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

Website Designer – Dana Busch

Production Assistant – Kristopher Cain

Teleprompter – Sue Willard-Kiess

Hair/Makeup - Amber Voner

Additional Footage

- Pond5/Burlingham
- Pond5/Kmotion
- Pond5/Digi_Dog

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.