

***AGAINST ALL ODDS***  
**EPISODE 22 – “SAMPLING DISTRIBUTIONS”**  
**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.

Meet the third grade class at Monica Ros School in Ojai, California – twenty rambunctious 9-year-olds who are about to become a living demonstration of how, with the right statistical tools, we can use samples from a population to make inferences about the population as a whole.

Remember, we get statistics from sample data, while parameters are generally unknown because they describe an entire population.

Here's our population, the nine-year olds at Monica Ros, lining up according to their height – 50-inchers at one end, 57-inchers at the other, with most of them clustered around the center near 53 or 54 inches. This is our population distribution – the distribution of values in our whole population of nine-year-olds in the class.

When we plot this out, you won't be surprised by now to see that the shape of the distribution is approximately a Normal curve, from which we can calculate the exact mean,  $\mu$  – 53.4 inches – and the standard deviation,  $\sigma$  – 1.8 inches.

Okay, now we are going to take some random samples from our population. Here comes the first sample of 4 students. Throughout this example we're going to stick with 4 as our sample size.

Once again we calculate the mean – in this case the sample mean, or "x-bar" – which comes out at 53 inches. Now we take a different random sample of 4. This time the sample mean is 52.25 inches. From a third random sample we get an "x-bar" of 52.75 inches. From a fourth, a different distribution but an "x-bar" that happens to be the same as the last. And from a fifth random sample we get a sample mean of 53.25 inches.

We can keep going until we've selected all possible samples of 4 from our population of 20.

And now we can plot out all the sample means we calculated. What we get is a distribution of our sample means – or to put it another way, a sampling distribution of "x-bar": a distribution of "x-bar" values from all possible samples of size 4. And like any distribution it can be described by its shape, center and spread. The shape is our familiar Normal curve. The center is the mean of "x-bar," 53.4 inches – which is exactly what we got for the mean when we calculated it from the population as a whole.

But when we directly compare the population distribution to our sampling distribution we can see that while the center is the same, the spread of the sampling distribution is much less. In fact, sample means are always less variable than individuals.

That makes sense. A random sample should include a variety of individuals – from our population, some short, some medium, some tall. The sample mean literally averages out that variety, so we see less variation.

There is a precise relationship between the standard deviation of the sample mean, and the standard deviation of the individual heights. Here it is. The standard deviation of the sample mean is simply the population standard deviation, divided by the square root of the sample size – in our case, 4. So the standard deviation of our sample mean is 0.9 inches.

Let's see how we can put this fact to use in a situation with more at stake than figuring out the average height of a third grade class – a manufacturing plant for circuit boards. Actually, this is a scene you don't see that much any more – an electronics manufacturing plant in the United States. Today, many electronic assembly plants are overseas, and involve components that are a lot smaller than these. But we are interested in how statistics helps control quality in manufacturing, and those principles haven't changed.

A key part of the manufacturing process is when the components on the board are connected together by passing it through a bath of molten solder. If things go wrong – say the temperature of the solder or the speed of the conveyer isn't right – then the connections on the boards will be faulty. Workers can't wait until the boards reach the end of the line to spot a problem that's occurring here, so after they've passed through the soldering bath, an inspector randomly selects boards for a quality check.

A score of 100 is the standard, with some a little higher, some a little lower – the natural variation inherent in any manufacturing process. The goal of the quality control process is to spot if this variation starts drifting out of the acceptable range, suggesting a problem with the soldering bath.

Here is the distribution of the quality scores, a Normal curve centered on 100. Its standard deviation, derived from the company's experience with the process, is 4.

So how does what we've learned about sampling distributions help quickly spot when things go awry? The inspector's random sampling of the boards has a sample size of 5. Let's see what happens when we take repeated samples.

From our first five boards, we get an "x-bar" of 99.4. From our second sample of five different boards we get a different "x-bar," 101.6. We keep on sampling,

calculating each “x-bar,” until we can start building a sampling distribution – which is, as expected, centered on the mean of the population, 100. But it has a smaller standard deviation of around 1.79, as you can see from the formula.

Here’s how this is useful in the statistical quality process known as the “x-bar” control chart. The inspector plots the values of “x-bar” against time. As expected, there’s variation. The goal of the process is to distinguish chance variation from the extra variation that shows something is going wrong.

Here’s the normal distribution of “x-bar.” If everything is going well, the mean will be at 100. Now, recall the 68-95-99.7 rule for any normal distribution. Almost all observations, 99.7 percent in fact, will lie within 3 standard deviations of the mean. Going out three standard deviations from the mean gives us what this quality control method calls control limits. As long as the quality scores remain normally distributed with mean 100 and standard deviation 4 – that is, as long as the soldering process continues its past pattern – almost all the “x-bar” points on the chart will fall between the control limits. The process is said to be “in control,” when its pattern of variation is stable over time.

A point outside the limit is evidence that the process has become more variable or that its mean has shifted; in other words, that it’s gone “out of control.” As soon as an inspector sees a point like this, it’s a signal to ask, what’s gone wrong?

So far we’ve been talking about population distributions that follow a roughly normal curve. What if we were to take a very different population distribution?

The Mayor’s 24 Hour Constituent Service Hotline in Boston answers hundreds of thousands of calls every year. The operators handle everything from simple requests to more complicated questions about city services.

**Justin Holmes**

Typically, people call to report things like potholes, or streetlights that might be out in their neighborhood.

**Tara Blumstein**

People put trash out too early, or people have trash out too late. Or people aren’t keeping their property clean.

**Janine Coppola**

They call from some street that...a major...four-lane street where they say there’s a turkey running down the middle of the street.

**Tara Blumstein**

There was one where a skunk got stuck in a revolving door.

**Jessica Obasohan**

I had one recently about a little boy who wanted to request to speak to a Park Ranger with a horse.

**Pardis Sabeti**

The thing about calls to this and most other call centers is that the length of the calls varies widely.

**Justin Holmes**

So, the average length of our calls on the 24 Hour Hotline is about a minute and a half. But, obviously, that can range.

**Pardis Sabeti**

Most calls are relatively brief, but a few can just go on and on. The Mayor's Hotline answered a total of 21,669 calls in one month. If we plot the length of each call, the variable, against the number of calls, we'd get a distribution sharply skewed to the right. You can see it on this density curve representing the month's call durations.

If we sampled from this population, would our sampling distribution also be skewed to the right? You might expect so. So let's find out, using different sample sizes, starting with samples of size 10.

Let's take our first sample of size 10. For this sample, the mean length of the calls is 98.7 seconds. We can start to build a histogram using the sample mean. Let's continue taking samples of size 10, each time finding the mean of the sample, and adding it to our histogram. This final histogram is based on forty different samples of size ten.

We can do the same thing with 40 different samples of size 20.

And finally we can do it again with 40 different samples with an  $n$  of 60.

Now let's compare our sampling distributions with the original population distribution for all calls to the Mayor's hotline. The spread of all of the sampling distributions is smaller than the spread of the population distribution. And you can see that the spread of the sampling distributions tightens still further as the  $n$  increases. Interestingly, while the smallest sample size, with an  $n$  of 10, is still right skewed, the asymmetry decreases as the sample size gets bigger. By the time  $n$  equals 60, we've lost the skew and the sampling distribution of the sample mean looks pretty much Normal. The Normal quantile plot of "x-bar" data from 40 samples of size 60 looks like a pretty straight line, underscoring that our data are Normal.

This happens because with larger samples we are less likely to get all big or all small numbers. We usually get a mix. Some "x-bars" will be above mu, some

below. So we get an approximately Normal shape for the sampling distribution, even though the parent population isn't Normal.

What we've uncovered here is one of the most powerful tools statisticians possess, called the Central Limit Theorem. This states that, regardless of the shape of the population distribution, the sampling distribution of the sample mean will be approximately Normal if the sample size is large enough. It's because of the Central Limit Theorem that statisticians can generalize from sample data to the larger population.

We'll be seeing how that can be useful in later modules about confidence intervals and significance tests. Stay tuned! I'm Pardis Sabeti for *Against All Odds*.

## PRODUCTION CREDITS

Host – Dr. Pardis Sabeti

Writer/Producer/Director – Maggie Villiger

Associate Producer – Katharine Duffy

Editor – Seth Bender

Director of Photography – Dan Lyons

Additional Camera – Noah Brookoff

Audio – Dan Casey

Sound Mix – Richard Bock

Animation – Jason Tierney

Title Animation — Jeremy Angier

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

Website Designer – Dana Busch

Production Assistant – Kristopher Cain

Teleprompter – Kelly Cronin

Hair/Makeup - Amber Voner

Additional Footage

- Coast Learning Systems

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

Copyright © 2014 Annenberg Learner

## 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](http://www.learner.org).