Data and Sampling Distributions

Population vs Sample

In the era of big data is sampling needed?

Predictive models are typically developed and piloted with samples.

Traditional statistics focused very much on the lefthand side, using theory based on strong assumptions about the population. Modern statistics has moved to the righthand side, where such assumptions are not needed.

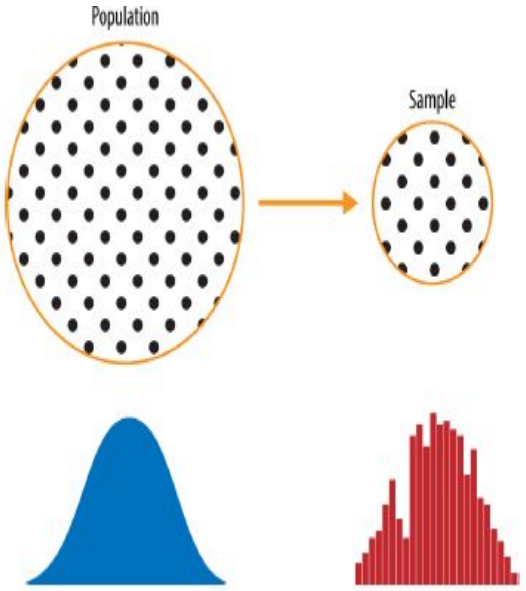


Figure 2-1. Population versus sample

Random Sampling and Sample Bias

A *sample* is a subset of data from a larger data set; statisticians call this larger data set the *population*

Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen for the sample at each draw. The sample that results is called a *simple random sample*.

- With Replacement: in which observations are put back in the population after each draw for possible future reselection.
- Without replacement :observations, once selected, are unavailable for future draws.

Data quality often matters more than data quantity if using sampling

Bias

Statistical bias refers to measurement or sampling errors that are systematic and produced by the measurement or sampling process

sample bias; that is, the sample was different in some meaningful and nonrandom way from the larger population it was meant to represent. Error due to random chance vs error due to bias

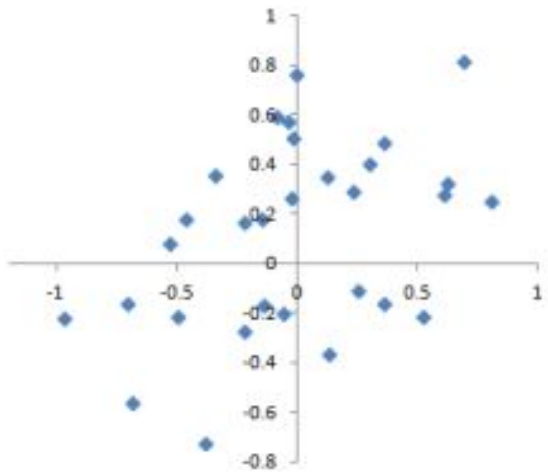


Figure 2-2. Scatterplot of shots from a gun with true aim

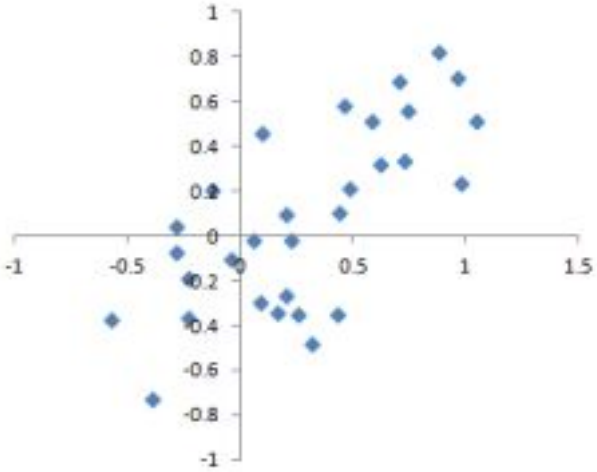


Figure 2-3. Scatterplot of shots from a gun with biased aim

The term nonrandom is important—hardly any sample, including random samples, will be exactly representative of the population.
Sample bias occurs when the difference is meaningful, and it can be expected to continue for other samples drawn in the same way as the first.

NOTE

Random Selection

To avoid the problem of sample bias.

Random sampling is not always easy. Proper definition of an accessible population is key.

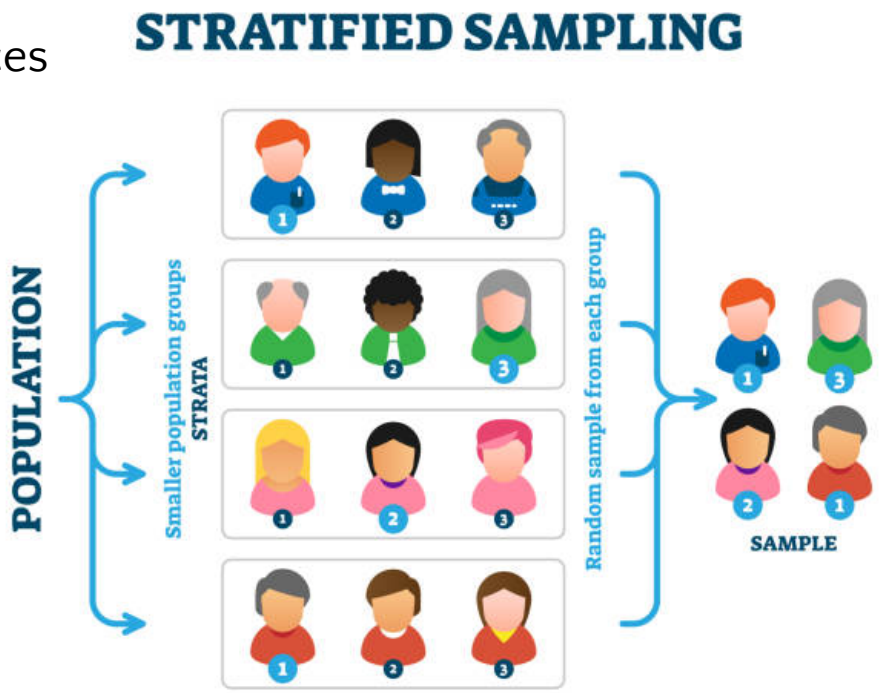
Reflection: Any Example?

Note:

- Bias occurs when measurements or observations are systematically in error because they are not representative of the full population.
- Data quality is often more important than data quantity, and random sampling can reduce bias and facilitate quality improvement that would otherwise be prohibitively expensive

Stratified sampling

Stratified sampling, the population is divided up into strata, and random samples are taken from each stratum.
Time and effort spent on random sampling not only reduces bias but also allows greater attention to data exploration and data quality



Sample Mean Versus Population Mean

The symbol x (pronounced “x-bar”) is used to represent the mean of a sample from a population, whereas μ is used to represent the mean of a population

Key-Terms

Key Terms for Random Sampling

Sample
A subset from a larger data set.

Population
The larger data set or idea of a data set.

N (n)
The size of the population (sample).

Random sampling
Drawing elements into a sample at random.

Stratified sampling
Dividing the population into strata and randomly sampling from each strata.

Stratum (pl., strata)
A homogeneous subgroup of a population with common characteristics.

Simple random sample
The sample that results from random sampling without stratifying the population.

Bias
Systematic error.

Sample bias
A sample that misrepresents the population.

Selection Bias: “If you torture the data long enough, sooner or later it will confess.”

Selection bias Bias resulting from the way in which observations are selected.

Data snooping Extensive hunting through data in search of something interesting.

Vast search effect Bias or no reproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.

Selection bias refers to the practice of selectively choosing data—consciously or unconsciously—in a way that leads to a conclusion that is misleading or ephemeral.

A form of selection bias of particular concern to data scientists is vast search effect.

If you repeatedly run different models and ask different questions with a large data set, you are bound to find something interesting. But is the result you found truly something interesting, or is it the chance outlier?

Regression to the Mean

successive measurements on a given variable.

Extreme observations tend to be followed by more central ones. Attaching special focus and meaning to the extreme value can lead to a form of selection bias

Sampling Distribution of a Statistic

sampling distribution of a statistic refers to the distribution of some sample statistic over many samples drawn from the same population

Sample statistic A metric calculated for a sample of data drawn from a larger population.

Data distribution The frequency distribution of individual values in a data set.

Sampling distribution The frequency distribution of a sample statistic over many samples or resamples.

Central limit theorem The tendency of the sampling distribution to take on a normal shape as sample size rises.

Standard error The variability (standard deviation) of a sample statistic over many samples (not to be confused with standard deviation, which by itself, refers to variability of individual data values).

Typically, a sample is drawn with the goal of measuring something (with a sample statistic) or modeling something (with a statistical or machine learning model). Since our estimate or model is based on a sample, it might be in error; it might be different if we were to draw a different sample. We are therefore interested in how different it might be – a key concern is sampling variability

Sampling Distribution of a Statistic

The distribution of a sample statistic such as the mean is likely to be more regular and bell-shaped than the distribution of the data itself. The larger the sample the statistic is based on, the more this is true. Also, the larger the sample, the narrower the distribution of the sample statistic.

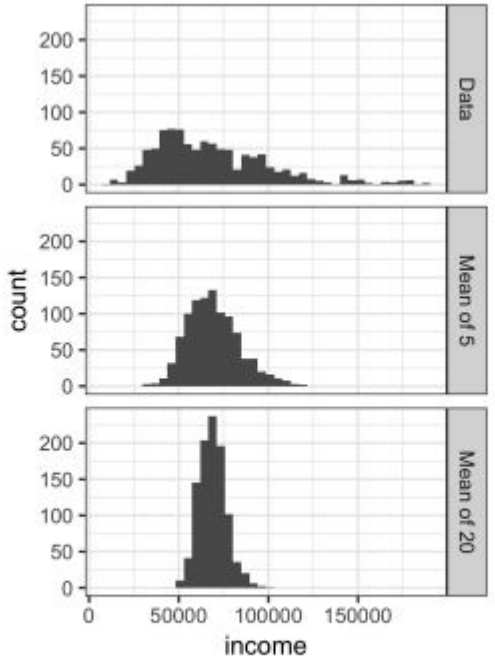


Figure 2-6. Histogram of annual incomes of 1,000 loan applicants (top), then 1,000 means of n=5 applicants (middle), and finally 1,000 means of n=20 applicants (bottom)

The histogram of the individual data values is broadly spread out and skewed toward higher values, as is to be expected with income data. The histograms of the means of 5 and 20 are increasingly compact and more bell-shaped.

Central Limit Theorem

The means drawn from multiple samples will resemble the familiar bell-shaped normal curve even if the source population is not normally distributed, provided that the sample size is large enough.

Conceptually, you can imagine the bootstrap as replicating the original sample thou- sands or millions of times so that you have a hypothetical population that embodies all the knowledge from your original sample (it's just larger)

We can also do sample with replacement.

Standard Error

The standard error is a single metric that sums up the variability in the sampling distribution for a statistic. The standard error can be estimated using a statistic based on the standard deviation s of the sample values, and the sample size n.

$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

Standard deviation (which measures the variability of individual data points)

standard error (which measures the variability of a sample metric).

Bootstarp

One easy and effective way to estimate the sampling distribution of a statistic, or of model parameters, is to draw additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample. This procedure is called the bootstrap, and it does not necessarily involve any assumptions about the data or the sample statistic being normally distributed.

The bootstrap can be used with multivariate data, where the rows are sampled as units

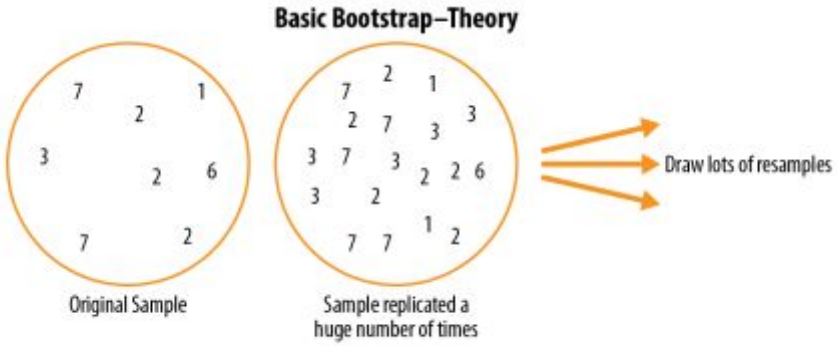


Figure 2-7. The idea of the bootstrap

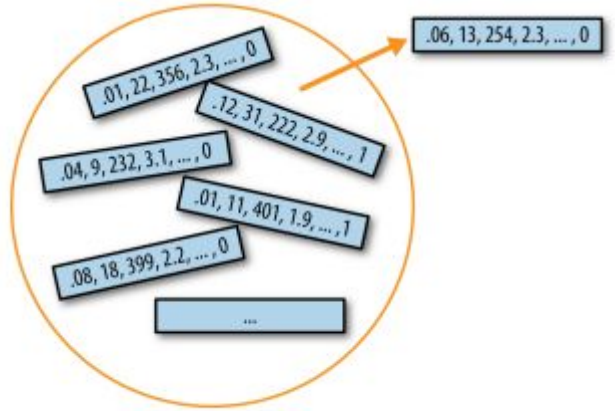


Figure 2-8. Multivariate bootstrap sampling

Boot strap

The bootstrap does not compensate for a small sample size; it does not create new data, nor does it fill in holes in an existing data set. It merely informs us about how lots of additional samples would behave when drawn from a population like our original sample.

The algorithm for a bootstrap resampling of the mean, for a sample of size n, is as follows:

1. Draw a sample value, record it, and then replace it.
2. Repeat n times.
3. Record the mean of the n resampled values.
4. Repeat steps 1–3 R times.
5. Use the R results to:

a. Calculate their standard deviation (this estimates sample mean standard error).

b. Produce a histogram or boxplot.

c. Find a confidence interval

Confidence Interval

Given a sample of size n, and a sample statistic of interest, the algorithm for a boot- strap confidence interval is as follows:

1. Draw a random sample of size n with replacement from the data (a resample).
2. Record the statistic of interest for the resample.
3. Repeat steps 1–2 many (R) times.
4. For an x% confidence interval, trim [(100-x) / 2]% of the R resample results from either end of the distribution. 5. The trim points are the endpoints of an x% bootstrap confidence interval.

Confidence Interval

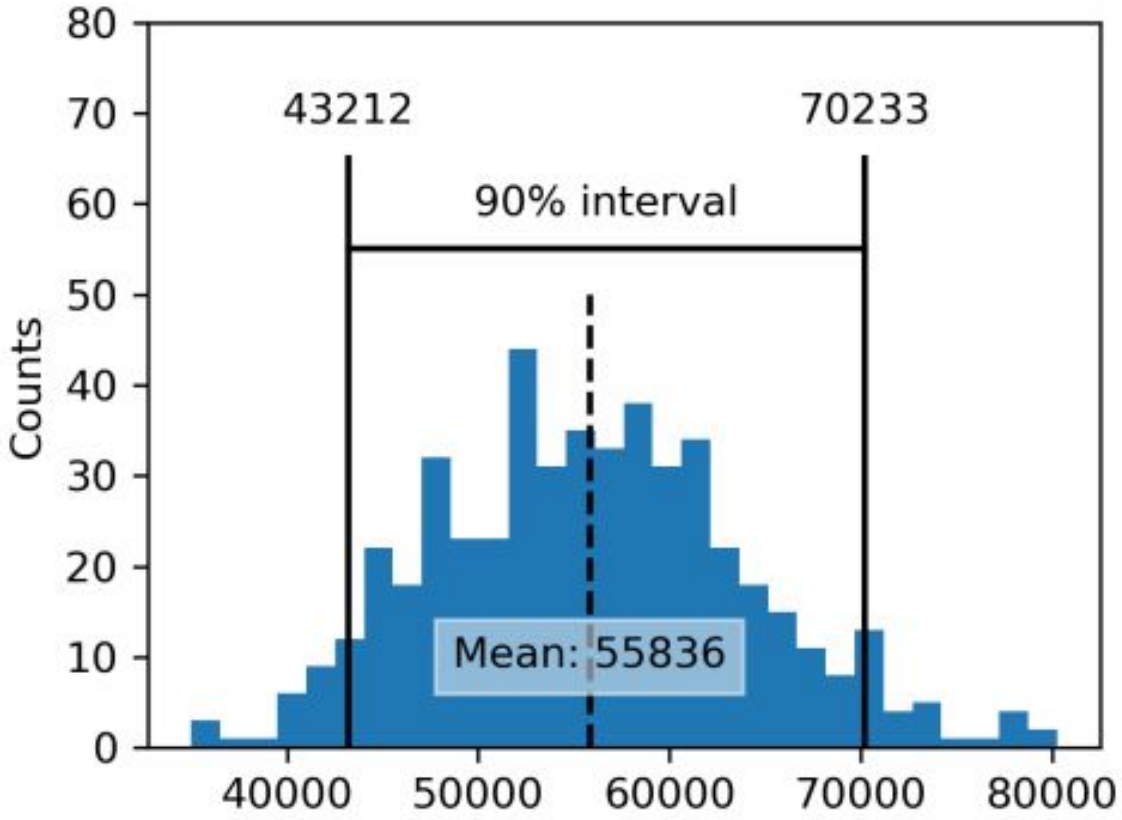


Figure 2-9. Bootstrap confidence interval for the annual income of loan applicants, based on a sample of 20

Normal Distribution

The bell-shaped normal distribution is iconic in traditional statistics.1 The fact that distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.

Error
The difference between a data point and a predicted or average value.

Standardize
Subtract the mean and divide by the standard deviation.

z-score
The result of standardizing an individual data point.

Standard normal
A normal distribution with mean = 0 and standard deviation = 1.

QQ-Plot
A plot to visualize how close a sample distribution is to a specified distribution, e.g., the normal distribution.

Standard Normal

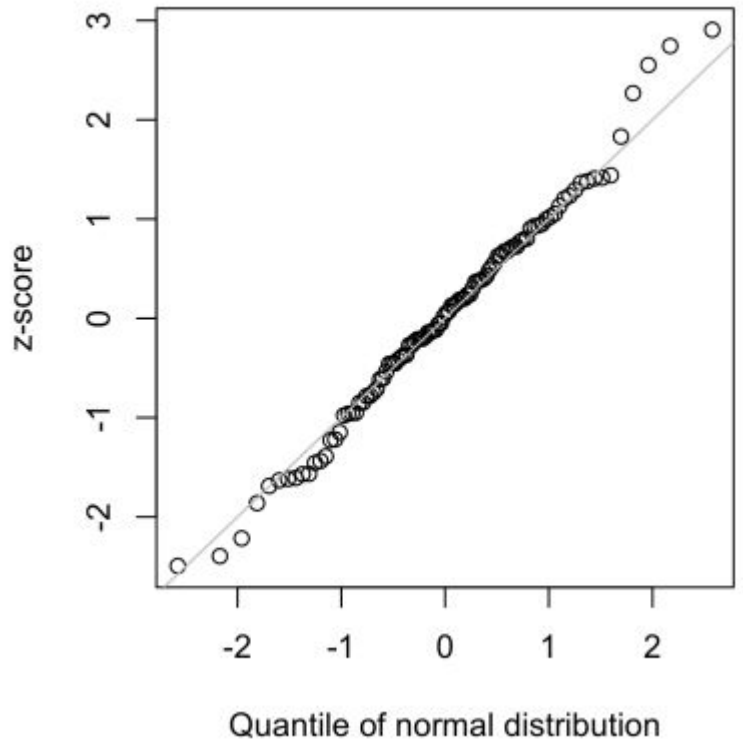
A standard normal distribution is one in which the units on the x-axis are expressed in terms of standard deviations away from the mean. To compare data to a standard normal distribution, you subtract the mean and then divide by the standard deviation; this is also called normalization or standardization.

The transformed value is termed a z-score, and the normal distribution is sometimes called the z-distribution.

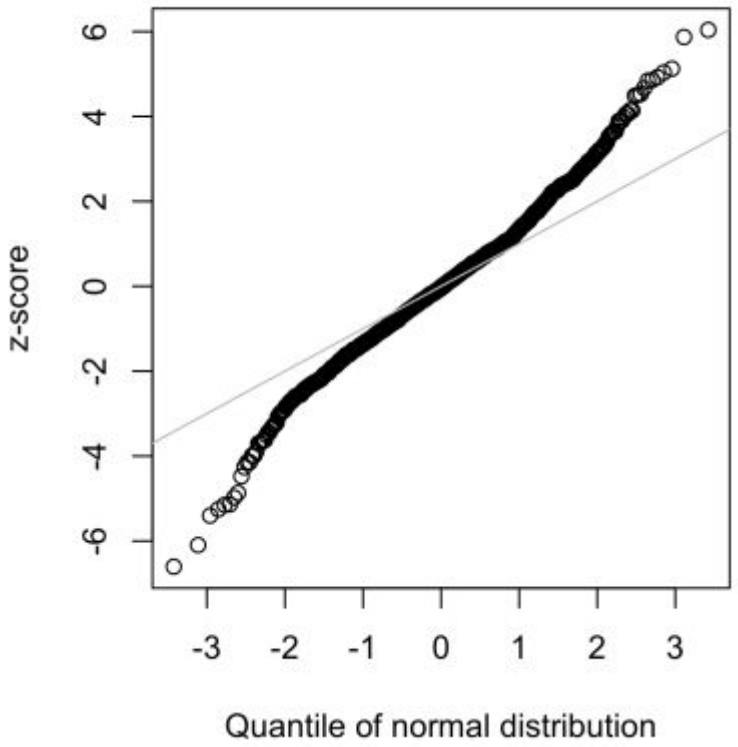
A QQ-Plot is used to visually determine how close a sample is to a specified distribution—in this case, the normal distribution. The QQ-Plot orders the z-scores from low to high and plots each value's z-score on the y-axis; the x-axis is the corresponding quantile of a normal distribution for that value's rank.

If the points roughly fall on the diagonal line, then the sample distribution can be considered close to normal

QQ Plot



Long-Tailed Distributions



Tail : The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.
Skew : Where one tail of a distribution is longer than the other.

• Assuming a normal distribution can lead to underestimation of extreme events (“black swans”)

Student's t-Distribution

The t-distribution is a normally shaped distribution, except that it is a bit thicker and longer on the tails. It is used extensively in depicting distributions of sample statistics. Distributions of sample means are typically shaped like a t-distribution, and there is a family of t-distributions that differ depending on how large the sample is. The larger the sample, the more normally shaped the t-distribution becomes.

Binomial Distribution

Yes/no (binomial) outcomes lie at the heart of analytics since they are often the cul- mination of a decision or other process; buy/don't buy, click/don't click, survive/die, and so on.

Central to understanding the binomial distribution is the idea of a set of trials, each trial having two possible outcomes with definite probabilities.

A binomial trial is an experiment with two possible outcomes: one with probability p and the other with probability 1 – p.

The binomial distribution is the frequency distribution of the number of successes (x) in a given number of trials (n) with specified probability (p) of success in each trial. There is a family of binomial distributions, depending on the values of n and p.

Chi-Square Distribution

An important idea in statistics is departure from expectation, especially with respect to category counts. Expectation is defined loosely as no correlation between variables or predictable patterns. This is also termed the “null hypothesis” or “null model”.

you might want to test whether one variable (gender) is independent of another (job)

The statistic that measures the extent to which results depart from the null expectation of independence is the chi-square statistic.

It is the difference between the observed and expected values, divided by the square root of the expected value, squared, then summed across all categories.

Poisson and Related Distributions

Many processes produce events randomly at a given overall rate—visitors arriving at a website, or cars arriving at a toll plaza (events spread over time); imperfections in a square meter of fabric, or typos per 100 lines of code (events spread over space).

Poisson Distributions

From prior aggregate data (for example, number of flu infections per year), we can estimate the average number of events per unit of time or space (e.g., infections per day, or per census unit). We might also want to know how different this might be from one unit of time/space to another.

The Poisson distribution tells us the distribu- tion of events per unit of time or space when we sample many such units.

Lambda

Lambda: This is the mean number of events that occurs in a specified interval of time or space.

A common technique is to generate random numbers from a Poisson distribution as part of a queuing simulation.

stats.poisson.rvs(2, size=100) This code will generate 100 random numbers from a Poisson distribution with $\lambda = 2$.

Exponential Distribution

Using the same parameter λ that we used in the Poisson distribution, we can also model the distribution of the time between events: time between visits to a website or between cars arriving at a toll plaza.

stats.expon.rvs (0.2, size=100) This code would generate 100 random numbers from an exponential distribution where the mean number of events per time period is 0.2.