Distributions

Stephen B. Holt, Ph.D.

Rockefeller College
of Public Affairs & Policy
University at Albany State University of New York

March 22, 2022

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
    - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
    - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
   - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
   - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
   - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
   - Pie and bar graphs - Depict the distribution of a categorical variable.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
    - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
    - Pie and bar graphs - Depict the distribution of a categorical variable.
    - Histogram - Depict the distribution of a quantitative variable.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
   - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
   - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
   - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
   - Pie and bar graphs - Depict the distribution of a categorical variable.
   - Histogram - Depict the distribution of a quantitative variable.
   - Scatterplot - Depict the relationship between two quantitative variables.

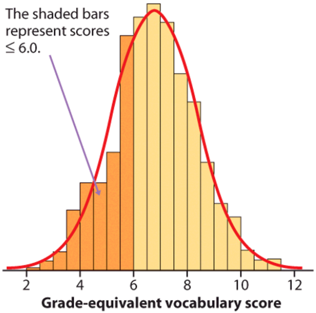
Basic Process

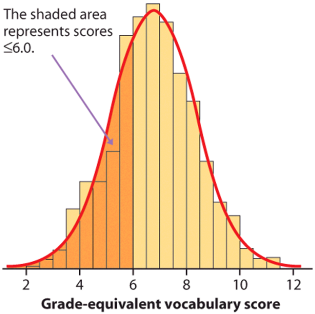
Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
    - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
    - Pie and bar graphs - Depict the distribution of a categorical variable.
    - Histogram - Depict the distribution of a quantitative variable.
    - Scatterplot - Depict the relationship between two quantitative variables.
    - Two-way table - Joint distribution of two categorical variables.

Density Curves and Distributions

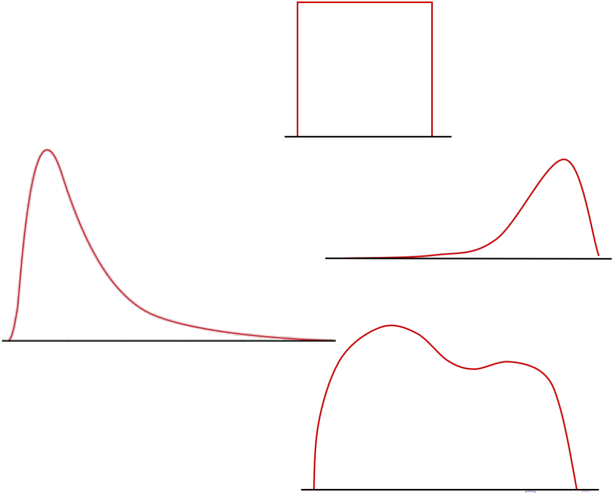
- Distributions describe the range and frequency of observations of a particular variable.
- While a histogram can document the number of observations at a given value of a variable, a density curve plots the area of a range of values the represents the proportion of all values within that range.
- The full area under a curve, consequently, is 100% of observations.





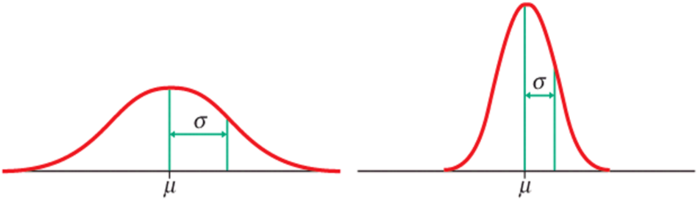
Density Curves and Distributions

Density curves can come in all shapes and sizes, some better understood than others.



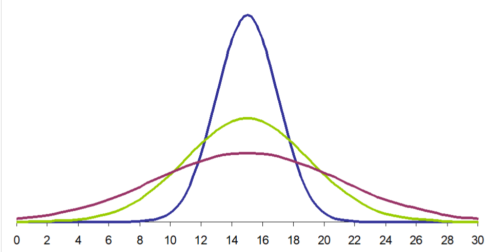
Normal Distributions

Normal – or Gaussian – distributions are a family of symmetrical, bell-shaped density curves defined by a mean μ (mu) and a standard deviation σ (sigma): $N(\mu, \sigma)$.

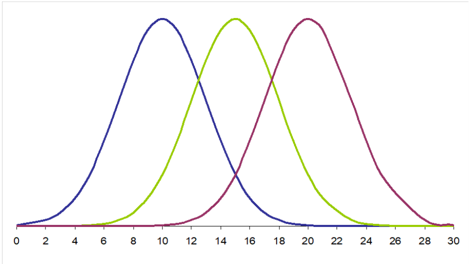


Normal Distributions

Distributions with the same mean and different spreads:

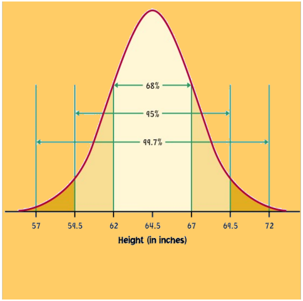


Distributions with different means and the same spreads:



Special Properties of Normal Distributions

- Normal distributions have mathematical properties that we use a lot in statistical inference. The 68-95-99.7 rule allows us to assess relative points in a distribution in terms of standard deviation. The rule is:
- 68% of observations fall within 1 standard deviation (σ) of the mean (μ)
- 95% of observations fall within 2 standard deviations (σ) of the mean (μ)
- 99.7% of observations fall within 3 standard deviations (σ) of the mean (μ)
- Note that μ and σ will always be used to refer to the (usually unknown) population mean and s.d., while $\overline{X}$ and s_x will always refer to a sample mean and s.d.



What are sampling distributions?

The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population. It is a theoretical idea — we do not actually build it.

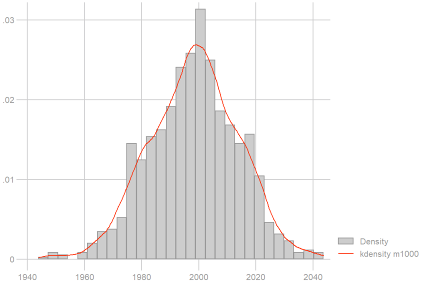
The sampling distribution of a statistic is the **probability distribution** of that statistic.

Sampling distribution of the sample mean

If we **were** to build the sampling distribution, we would take many, many samples of a given sample size (n) from a population with mean μ and standard deviation σ.
Some of these samples will have a mean above the population mean μ and some will be below. If we plot the mean of each sample in a distribution, we will have the sampling distribution of μ. Note, this is different than a distribution of x from a single sample.

Say we have a population of 80,000 people and $\mu = 2000$ sq. ft. houses. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985$
2. SRS size 1000 $\rightarrow \overline{x} = 1999$
3. SRS size 1000 $\rightarrow \overline{x} = 2010$



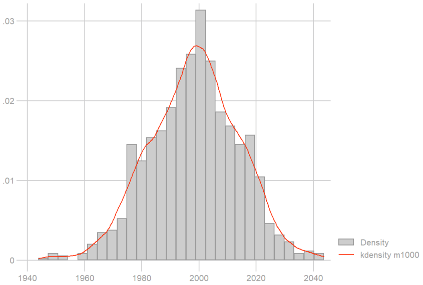
For any population with mean μ and standard deviation σ:

- The mean (or center) of the sampling distribution of $\overline{x}$ is equal to the population mean μ (the **law of large numbers**)
- The standard deviation of the sampling distribution is $\dfrac{\sigma}{\sqrt{n}}$, where n is the sample size (the **central limit theorem**).

These two properties of the sampling distribution are dictated by the **law of large numbers** and the **central limit theorem**.

Say we have a population of 80,000 people and $\mu = 2000$ sq. ft. houses. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985$
2. SRS size 1000 $\rightarrow \overline{x} = 1999$
3. SRS size 1000 $\rightarrow \overline{x} = 2010$



Law of Large Numbers

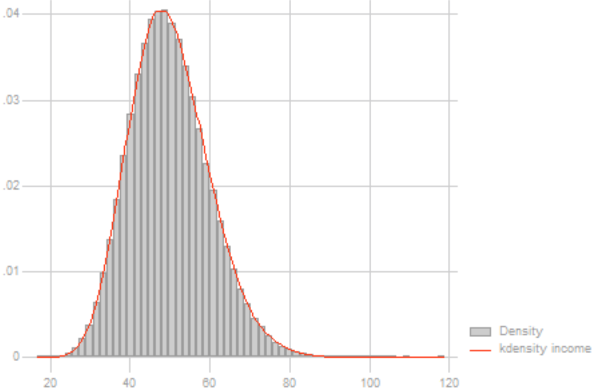
Here, I will use data from a simulation to demonstrate the law of large numbers and central limit theorem. The **law of large numbers** dictates that as a sample size gets larger and larger, the average of the sample will be the same as the average of the population.

In other words, the larger the sample we collect information from, the more likely the samples will have roughly the same average. This is referred to as an estimate's **consistency**.

Since sampling distributions are theoretically built with infinite samples, or very large numbers of samples of the same variable from the same population, the law of large numbers dictates that the mean of their distribution will be equal to the mean of the population.

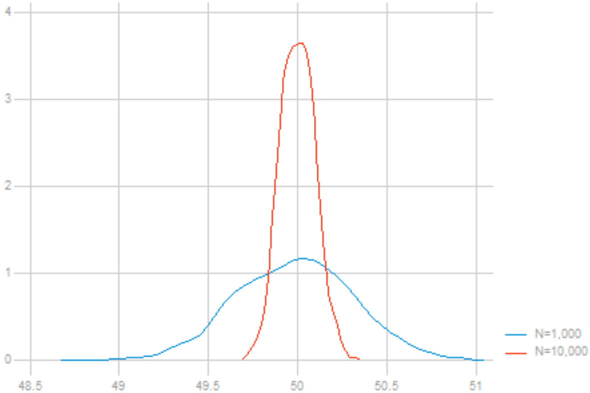
Example of Law of Large Numbers

I created a fake population of 800,000 people and data on their incomes (in \$1000s) with a mean of \$50K. Unlike most real-world situations, this means we **know** the true value of μ and σ. In this case, $\mu = 49.99$ and $\sigma = 9.99$. Below is the distribution from the population:



Example of Law of Large Numbers

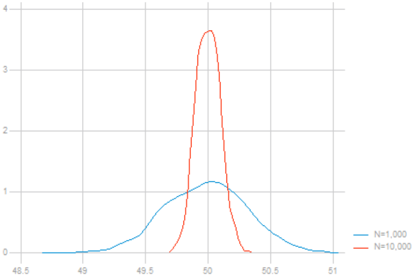
I simulated taking a thousand random samples with sample sizes of 1,000 observations and 10,000 observations and calculated the average of each sample. The distributions below show the sample means of all 1,000 samples for each sample size:



Example of Law of Large Numbers

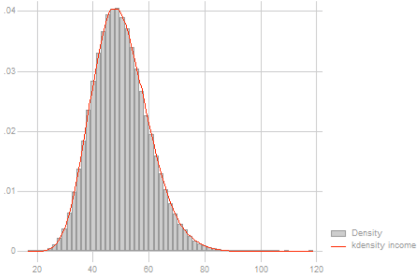
A few things to note about what the simulation is showing:

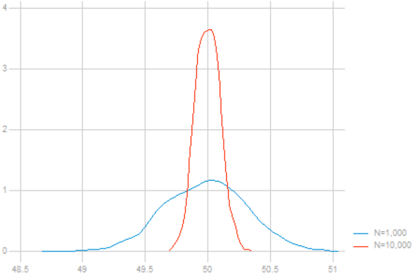
1. The range for the population is from 20 to 120 while the sample distributions both range from 1.3 below to 1.01 above the μ of 50 (with a smaller range for the larger sample).
2. The high peak for samples of 10,000 people shows that a much larger proportion of samples have an estimated mean ($\overline{x}$) equal to the underlying population mean (μ).
3. Larger sample sizes yield means that converge closer and closer to the underlying population average.



Example of Central Limit Theorem

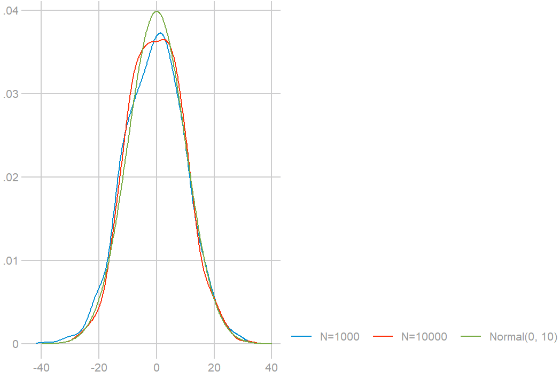
The central limit theorem provides the basis for estimating standard deviations from samples and performing hypothesis tests. The central limit theorem states that the sampling distribution of means will always be approximately normally distributed, even when the variable is not normally distributed in the population. Notice that our population in the previous example had a long tail to the right, but the sampling distributions did not:





Example of Central Limit Theorem

Rescaling and recentering both distributions shows that the sampling distribution for both 1,000 observation samples and 10,000 observations samples is nearly identical to a normal distribution.



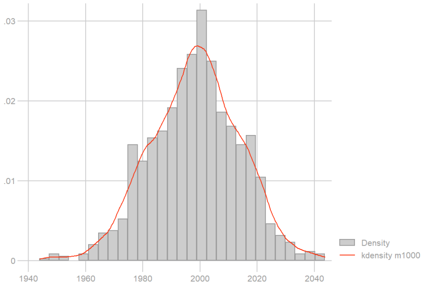
Statistical Confidence

Although the sample mean is a unique number for any particular sample, if you pick a different sample you will probably get a different sample mean.
In fact, you could get many different values for the sample mean, and virtually none of them would actually equal the true population mean, μ.
But due to the Law of Large Numbers, we know it will be close, particularly when we take larger samples.

$\mu = ?$; $\sigma = 500$. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985$
2. SRS size 1000 $\rightarrow \overline{x} = 1999$
3. SRS size 1000 $\rightarrow \overline{x} = 2010$

Sampling distribution of $\overline{x}$



Confidence Intervals

We can use our knowledge that the sampling distribution will be 1) centered around the population mean because of the law of large numbers and 2) will have a normal distribution because of the central limit theorem to help us assess the quality and confidence we have that our sample average accurately represents the population average. We can do this using the standard deviation of the sampling distribution to calculate what's known as a **confidence interval**.

The **confidence interval** is a range of values with an associated probability or confidence level C. The probability quantifies the chance that the interval contains the true population parameter.

$\mu = ?$; $\sigma = 500$; $\sigma_{\overline{x}} = \frac{500}{\sqrt{1000}} = 15.81$. We draw a bunch of simple random samples (SRS) and calculate the mean sq. ft.:

1. SRS size 1000 $\rightarrow \overline{x} = 1985 \rightarrow \overline{x} \pm 31.62 \rightarrow 1985 \pm 31.62$
2. SRS size 1000 $\rightarrow \overline{x} = 1999 \rightarrow \overline{x} \pm 31.62 \rightarrow 1999 \pm 31.62$
3. SRS size 1000 $\rightarrow \overline{x} = 2010 \rightarrow \overline{x} \pm 31.62 \rightarrow 2010 \pm 31.62$

Confidence Intervals

A confidence interval can be calculated as $\overline{x} \pm m$.
m refers to the **margin of error** or the z-score the analyst is choosing for C confidence levels. Thus, $m = z * \frac{\sigma}{\sqrt{n}}$.

As an example, mean of 120 and margin of error of 6:
$120 \pm 6 \rightarrow$ confidence interval ranges from 114 to 126.

You interpret this as: 95% of sample means of samples of n size will be between 114 and 126 or I am 95% confident that the true population mean is between 114 and 126.

A confidence level C (in %) indicates the probability that the μ falls within the interval. It represents the area under the normal curve within $\pm m$ of the center of the curve.

Link between Confidence Level and Margin of Error

The confidence level C determines the value of z*.
The margin of error also depends on z*.

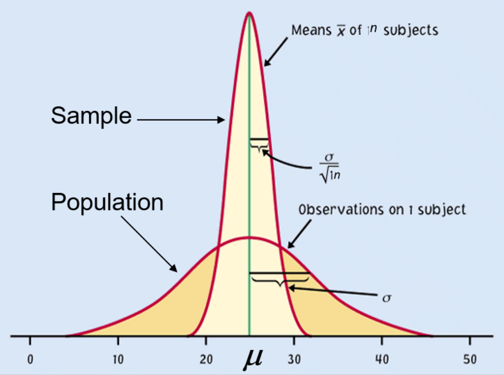
The C is the percent of the distribution you would like to fall between the two end points you plan to calculate. That percent provides a measure of how likely you are to have an estimated average that is unrealistic or far from the true average. Once you've decided on the percentage for C, you choose the z-score associated with C and use the formula $m = z * \frac{\sigma}{\sqrt{n}}$ to calculate the end points of the confidence interval.

A higher C means more confidence that the true mean is within the interval, but it also means a wider interval (less precise estimates).

A lower C means less confidence in our estimate, but a narrower interval (more precise estimates).
We generally prefer to be cautious. A rule of thumb is to use 95% confidence intervals (i.e., a z-score of 2).

Implications

We *don't need* to take endless random samples to "rebuild" the sampling distribution and find μ at its center. We only need one SRS of size n and we can use the properties of the sampling distribution of means to infer the population mean μ. The central limit theorem and law of large numbers let us have pre-existing knowledge of the sampling distribution without building it from scratch each time.


Means $\bar{x}$ of ın subjects
Sample
$\frac{\sigma}{\sqrt{1n}}$
Population
Observations on ı subject
σ
μ