Distributions

Stephen B. Holt, Ph.D.

ROCKEFELLER COLLEGE OF PUBLIC AFFAIRS & POLICY
UNIVERSITY AT ALBANY State University of New York

March 1, 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.

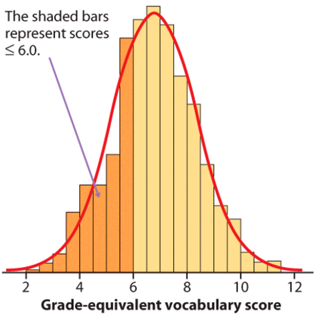
Basic Process

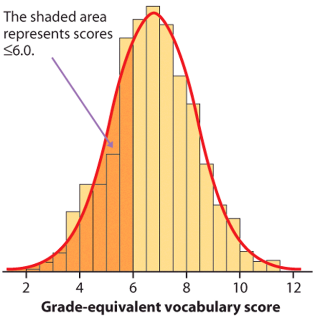
Most policy research involves deceptively simple steps:

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

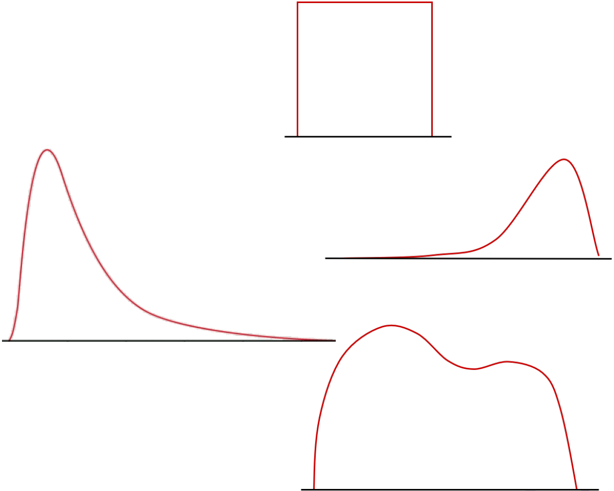
- Distributions describe the range and frequency of observations of a particular variable.
- While a histogram can document the number of observations at a given value of a variable, a density curve plots the area of a range of values the represents the proportion of all values within that range.
- The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





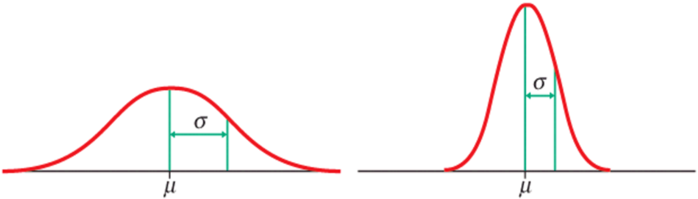
Density Curves and Distributions

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



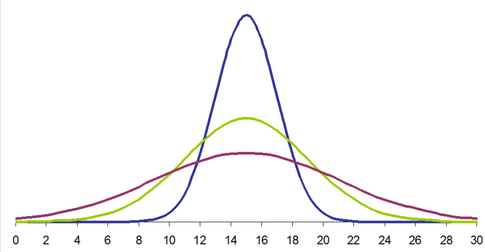
Normal Distributions

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): $N(\mu, \sigma)$.

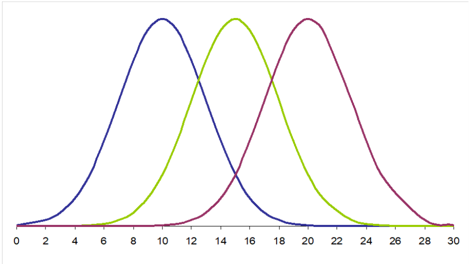


Normal Distributions

Distributions with the same mean and different spreads:

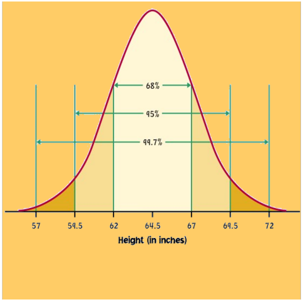


Distributions with different means and the same spreads:



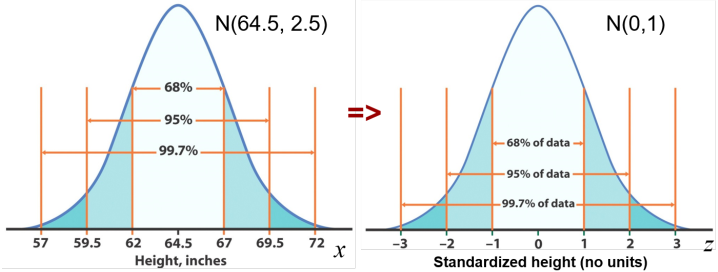
Special Properties of Normal Distributions

- 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.



Standard Normal Distributions

Since all normal distributions share the same properties, we can also express any normally distributed variable in terms of standard deviations. To do this, we would standardize our data, transforming the normal curve of $N(\overline{X}, S_x)$ to a standard normal curve of $N(0, 1)$. We do this by calculating the z-score for each observation of X.



Calculating Z-Scores

Calculating Z-scores is straightforward. The intent is to express X is terms of deviations from the mean using a standardized measure (in this case, s.d.). Thus, we use the z-score formula to calculate the number of standard deviations each value of X is from the mean.

$$Z = \frac{(X - \mu)}{\sigma} \tag{1}$$

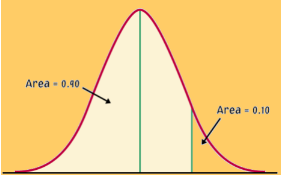
Note that when X is larger than μ, Z is positive; when X is smaller than μ, Z is negative.

Example: Heights

Women's height follows the $N(64.5'', 2.5'')$ distribution. What percent of women are shorter than 67 inches (5 ft. 6 inches)?
Mean: 64.5
S.D.: 2.5"
X: 67"
Let's calculate the z-score for X:

$$Z = \frac{(X - \mu)}{\sigma} \rightarrow Z = \frac{(67 - 64.5)}{2.5} \rightarrow \frac{2.5}{2.5} \rightarrow Z = 1. \tag{2}$$

We can use the 68-95-99.7 rule to deduce that the percent of women shorter than 67" should be about (.68 + half of 1-0.68) or .84 or 84% of women.



Z-Tables

- We can use a z-table to be more precise. A z-table allows us to examine the proportion of a distribution below a given z-score. Subtracting that proportion from 1 also gives us the proportion above a z-score.
- Here, we see that for a z of 1, we get an area of 0.8413 or 84.13% below.
- Subtracting from 1, we can see that 15.85% of women are **taller** than 67".

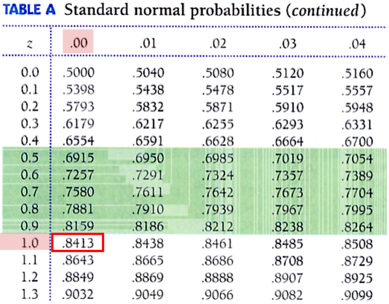


TABLE A Standard normal probabilities (*continued*)

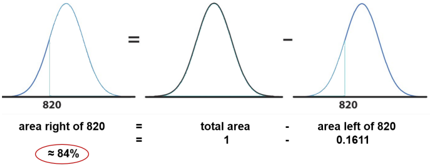
z	.00	.01	.02	.03	.04
0.0	.5000	.5040	.5080	.5120	.5160
0.1	.5398	.5438	.5478	.5517	.5557
0.2	.5793	.5832	.5871	.5910	.5948
0.3	.6179	.6217	.6255	.6293	.6331
0.4	.6554	.6591	.6628	.6664	.6700
0.5	.6915	.6950	.6985	.7019	.7054
0.6	.7257	.7291	.7324	.7357	.7389
0.7	.7580	.7611	.7642	.7673	.7704
0.8	.7881	.7910	.7939	.7967	.7995
0.9	.8159	.8186	.8212	.8238	.8264
1.0	.8413	.8438	.8461	.8485	.8508
1.1	.8643	.8665	.8686	.8708	.8729
1.2	.8849	.8869	.8888	.8907	.8925
1.3	.9032	.9049	.9066	.9082	.9099

Example 2 - Athletes and SATs

The National Collegiate Athletic Association (NCAA) requires Division I athletes to score at least 820 on the combined math and verbal SAT exam to compete in their first college year. The SAT scores of 2003 were approximately normal with mean 1026 and standard deviation 209.
What proportion of all students would be NCAA qualifiers (SAT $\geq$ 820)?
Mean: 1026
S.D.: 209
X: 820

$$Z = \frac{(X-\mu)}{\sigma} \rightarrow Z = \frac{(820-1026)}{209} \rightarrow \frac{-206}{209} \rightarrow Z = -0.99. \qquad (3)$$

A Z-table shows the area under -0.99 is 0.1611 or 1-16.11% of students would be qualifiers (83.89%).

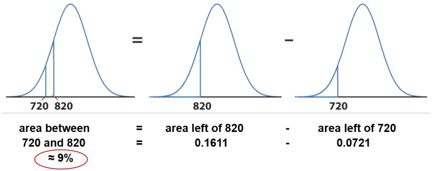


Example 3 - Athletes and SATs

The NCAA defines a “partial qualifier” eligible to practice and receive an athletic scholarship, but not to compete, with a combined SAT score of at least 720.
What proportion of all students who take the SAT would be partial qualifiers? That is, what proportion have scores between 720 and 820?
Mean: 1026
S.D.: 209
X: 720

$$Z = \frac{(X - \mu)}{\sigma} \rightarrow Z = \frac{(720 - 1026)}{209} \rightarrow \frac{-306}{209} \rightarrow Z = -1.46. \qquad (4)$$

A Z-table shows the area under -1.46 is 0.0721 or 7% of students. $16 - 7 =$ 9% of students have scores between 720 and 820.



Z-Tables

We may also want to find the observed range of values that correspond to a given proportion/ area under the curve. For that, we use Table A backward:

1. we first find the desired area/ proportion in the body of the table,
2. we then read the corresponding z-value from the left column and top row.

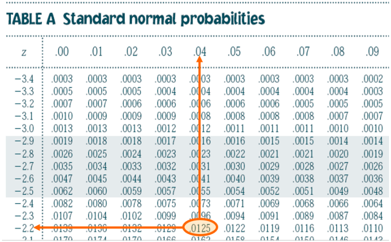


TABLE A Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143

For an area to the left of 1.25 % (0.0125), the z-value is -2.24

Attendance