Midterm Review

Stephen B. Holt, Ph.D.

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

March 8, 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.

Measures of Central Tendency

- Mean, average

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

  - Calculating in Stata: `sum varlist`

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

  - Calculating in Stata: `sum varlist`
- Median

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

  - Calculating in Stata: `sum varlist`
- Median
  - Interpretation: The exact center of an ordered distribution of a variable. From this point, 50% of observations have a higher value and 50% have a lower value.

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

  - Calculating in Stata: `sum varlist`
- Median
  - Interpretation: The exact center of an ordered distribution of a variable. From this point, 50% of observations have a higher value and 50% have a lower value.
  - Formula: 1. sort the data from lowest to highest; 2. if n is **odd**, the median is observation $(n+1)/2$ down the list; 3. if n is **even**, the median is the mean of the middle two observations.

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

  - Calculating in Stata: `sum varlist`
- Median
  - Interpretation: The exact center of an ordered distribution of a variable. From this point, 50% of observations have a higher value and 50% have a lower value.
  - Formula: 1. sort the data from lowest to highest; 2. if n is **odd**, the median is observation $(n+1)/2$ down the list; 3. if n is **even**, the median is the mean of the middle two observations.
  - Calculating in Stata: `sum varlist, detail`

Measures of Central Tendency

- Mean, average
  - Interpretation: Approximates a representative value of a variable from a population.
  - Formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

  - Calculating in Stata: `sum varlist`
- Median
  - Interpretation: The exact center of an ordered distribution of a variable. From this point, 50% of observations have a higher value and 50% have a lower value.
  - Formula: 1. sort the data from lowest to highest; 2. if n is **odd**, the median is observation $(n+1)/2$ down the list; 3. if n is **even**, the median is the mean of the middle two observations.
  - Calculating in Stata: `sum varlist, detail`
- Distance between median and mean suggests size and direction of skew.

Quartiles

- Quartiles

Quartiles

- Quartiles
  - Interpretation: The exact center above and below the median value. A large value distance suggests a wide spread and vice versa.

Quartiles

- Quartiles
  - Interpretation: The exact center above and below the median value. A large value distance suggests a wide spread and vice versa.
  - Formula: After finding the median, find the median of the observations above the median and then find the median of the observations below the median. In both cases, exclude the median.

Quartiles

- Quartiles
  - Interpretation: The exact center above and below the median value. A large value distance suggests a wide spread and vice versa.
  - Formula: After finding the median, find the median of the observations above the median and then find the median of the observations below the median. In both cases, exclude the median.
  - Calculating in Stata: `sum varlist, detail` (25% and 75% give you the values of the quartiles; 50% is, of course, the median)

Inter-quartile Range (IQR) and outliers

1. Subtract bottom quartile value from top quartile value ($IQR = TopQ - BottomQ$).
2. Find the outlier distance by multiplying 1.5 times the IQR ($OD = 1.5 \times IQR$)
3. Find the outlier thresholds by adding the outlier distance to the top quartile and substracting the outlier distance from the bottom quartile. Values above and below these thresholds, respectively, are outliers.

Standard Deviations

- Standard Deviation

Standard Deviations

- Standard Deviation
  - Interpretation: A standardized unit that measures distance relative to the mean. Large standard deviations means a high spread and vice versa.

Standard Deviations

- Standard Deviation
  - Interpretation: A standardized unit that measures distance relative to the mean. Large standard deviations means a high spread and vice versa.
  - Formula:

$$s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2 \tag{2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2} \tag{3}$$

Standard Deviations

- Standard Deviation
  - Interpretation: A standardized unit that measures distance relative to the mean. Large standard deviations means a high spread and vice versa.
  - Formula:

$$s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2 \tag{2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2} \tag{3}$$

  - Calculating in Stata: `sum varlist`

Z-Scores

- Z-Scores

Z-Scores

- Z-Scores
  - Interpretation: A standardized measure of an observation's distance from the sample mean expressed in terms of standard deviations.

Z-Scores

- Z-Scores
  - Interpretation: A standardized measure of an observation's distance from the sample mean expressed in terms of standard deviations.
  - Formula:

$$Z = \frac{(X - \mu)}{\sigma} \tag{4}$$

Z-Scores

- Z-Scores
  - Interpretation: A standardized measure of an observation's distance from the sample mean expressed in terms of standard deviations.
  - Formula:

$$Z = \frac{(X - \mu)}{\sigma} \tag{4}$$

  - Calculating in Stata: 1. `egen mean_x = mean(varx)`; 2. `egen sd_x = sd(varx)` 3. `gen z_x = ((varx - mean_x)/sd_x)`

Z-Scores

- Z-Scores
  - Interpretation: A standardized measure of an observation's distance from the sample mean expressed in terms of standard deviations.
  - Formula:

$$Z = \frac{(X - \mu)}{\sigma} \tag{4}$$

  - Calculating in Stata: 1. `egen mean_x = mean(varx)`; 2. `egen sd_x = sd(varx)` 3. `gen z_x = ((varx - mean_x)/sd_x)`
  - 68-95-99.7 rule: 68% of observations are between -1 and 1 s.d.'s; 95% of observations are between -2 and 2 s.d.'s; 99.7% of observations are between -3 and 3 s.d.'s.

Using Z-Tables

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Pearson's R

- Pearson's R Coefficient

Pearson's R

- Pearson's R Coefficient
  - Interpretation: Standardized measure of the strength of the relationship between two variables, ranging from -1 to 1. Closer to -1 and 1 represents strong relationships.

Pearson's R

- Pearson's R Coefficient
  - Interpretation: Standardized measure of the strength of the relationship between two variables, ranging from -1 to 1. Closer to -1 and 1 represents strong relationships.
  - Formula:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) \tag{5}$$

Pearson's R

- Pearson's R Coefficient
  - Interpretation: Standardized measure of the strength of the relationship between two variables, ranging from -1 to 1. Closer to -1 and 1 represents strong relationships.
  - Formula:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) \tag{5}$$

  - Calculating in Stata: `corr var1 var2`

Categorical Variables

- Relationships of Categorical Variables

Categorical Variables

- Relationships of Categorical Variables
  - Interpretation: The relationship between categorical variables is expressed in proportions or probabilities. Conditional on being in a category in variable A, the proportion of observations in a category in variable B provides the conditional probability.

Categorical Variables

- Relationships of Categorical Variables
  - Interpretation: The relationship between categorical variables is expressed in proportions or probabilities. Conditional on being in a category in variable A, the proportion of observations in a category in variable B provides the conditional probability.
  - Formula:

$$P = \frac{N_{cell}}{N_{column}} or \frac{N_{cell}}{N_{row}} \tag{6}$$

Categorical Variables

- Relationships of Categorical Variables
  - Interpretation: The relationship between categorical variables is expressed in proportions or probabilities. Conditional on being in a category in variable A, the proportion of observations in a category in variable B provides the conditional probability.
  - Formula:

$$P = \frac{N_{cell}}{N_{column}} or \frac{N_{cell}}{N_{row}} \tag{6}$$

  - Calculating in Stata: `tab var1 var2, col row` (for a two-way table)

Categorical Variables

- Relationships of Categorical Variables
  - Interpretation: The relationship between categorical variables is expressed in proportions or probabilities. Conditional on being in a category in variable A, the proportion of observations in a category in variable B provides the conditional probability.
  - Formula:

$$P = \frac{N_{cell}}{N_{column}} or \frac{N_{cell}}{N_{row}} \tag{6}$$

  - Calculating in Stata: `tab var1 var2, col row` (for a two-way table)
  - Two-way tables provide the marginal distribution (the proportion of each category for the whole sample in both the row variable and column variable) and the conditional distributions of both variables (the row and column percents in each cell or ex. how variable A is distributed conditional on being in category 1 of variable B)

Study Designs

Common Study Designs

Study Designs

Common Study Designs

- Observational:

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship
- Quasi-experimental:

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship
- Quasi-experimental:
  - Use advanced statistical techniques to estimate effects of treatments on people in the real world

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship
- Quasi-experimental:
  - Use advanced statistical techniques to estimate effects of treatments on people in the real world
  - Good for identifying a causal link between an intervention and an outcome.

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship
- Quasi-experimental:
  - Use advanced statistical techniques to estimate effects of treatments on people in the real world
  - Good for identifying a causal link between an intervention and an outcome.
- Experimental:

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship
- Quasi-experimental:
  - Use advanced statistical techniques to estimate effects of treatments on people in the real world
  - Good for identifying a causal link between an intervention and an outcome.
- Experimental:
  - Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.

Study Designs

Common Study Designs

- Observational:
  - Record data on individuals without attempting to influence the responses
  - Good for describing a trend or theoretically important relationship
- Quasi-experimental:
  - Use advanced statistical techniques to estimate effects of treatments on people in the real world
  - Good for identifying a causal link between an intervention and an outcome.
- Experimental:
  - Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.
  - Good for identifying a causal link between an intervention and an outcome.

Sampling

Common sample designs:

Sampling

Common sample designs:

- Convenience sampling:

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling:

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling:
  - Individuals are randomly selected. Each individual in the population has the same probability of being in the sample. Example: Public opinion polls.

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling:
  - Individuals are randomly selected. Each individual in the population has the same probability of being in the sample. Example: Public opinion polls.
- Stratified random sample:

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling:
  - Individuals are randomly selected. Each individual in the population has the same probability of being in the sample. Example: Public opinion polls.
- Stratified random sample:
  - a series of random sampling performed on subgroups of a given population. Examples: Some government surveys.

Sampling

Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling:
  - Individuals are randomly selected. Each individual in the population has the same probability of being in the sample. Example: Public opinion polls.
- Stratified random sample:
  - a series of random sampling performed on subgroups of a given population. Examples: Some government surveys.
- Multiple stage random sample:

Sampling

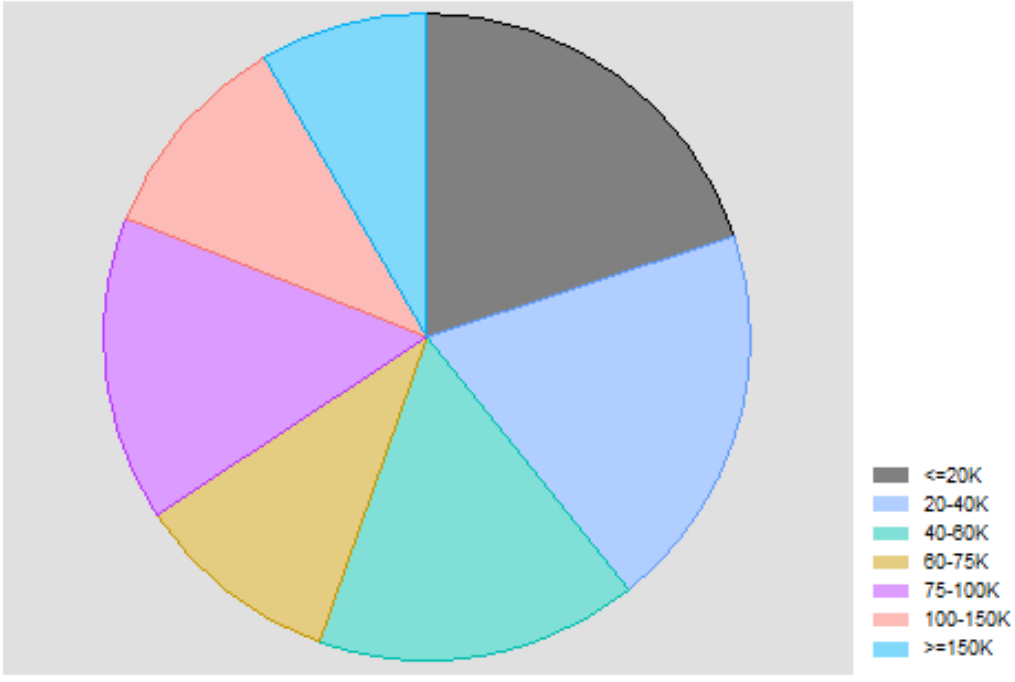
Common sample designs:

- Convenience sampling:
  - Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling:
  - Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling:
  - Individuals are randomly selected. Each individual in the population has the same probability of being in the sample. Example: Public opinion polls.
- Stratified random sample:
  - a series of random sampling performed on subgroups of a given population. Examples: Some government surveys.
- Multiple stage random sample:
  - select groups within a population in stages, resulting in a sample consisting of clusters of individuals. Examples: Many government studies.

Variable Types

- Response, dependent variables - the variable that measures the outcome being studied (e.g., student learning, physical health, etc.)
- Explanatory, independent variables - the variable that measures the factor or treatment believed to be related to changes in the outcome of interest in a study
- Lurking, omitted variable - the variable not accounted for in a study design that might explain all or part of an observed relationship

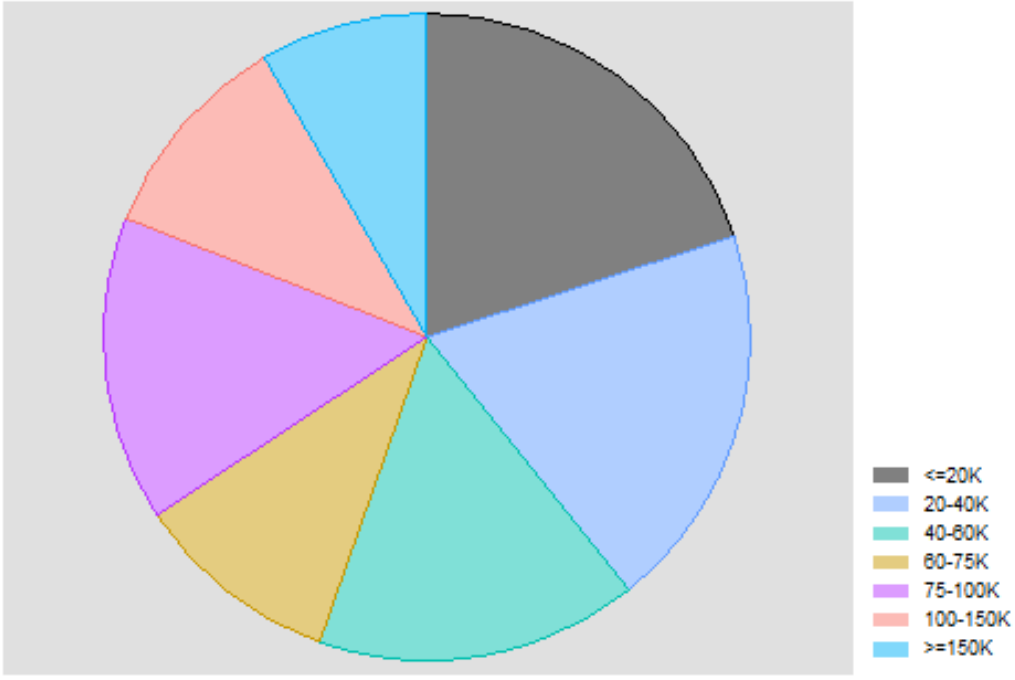
Pie Graphs



Code: `graph pie tucaseidr, over(hhincome)`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

- Requires a numeric variables that identifies observations.

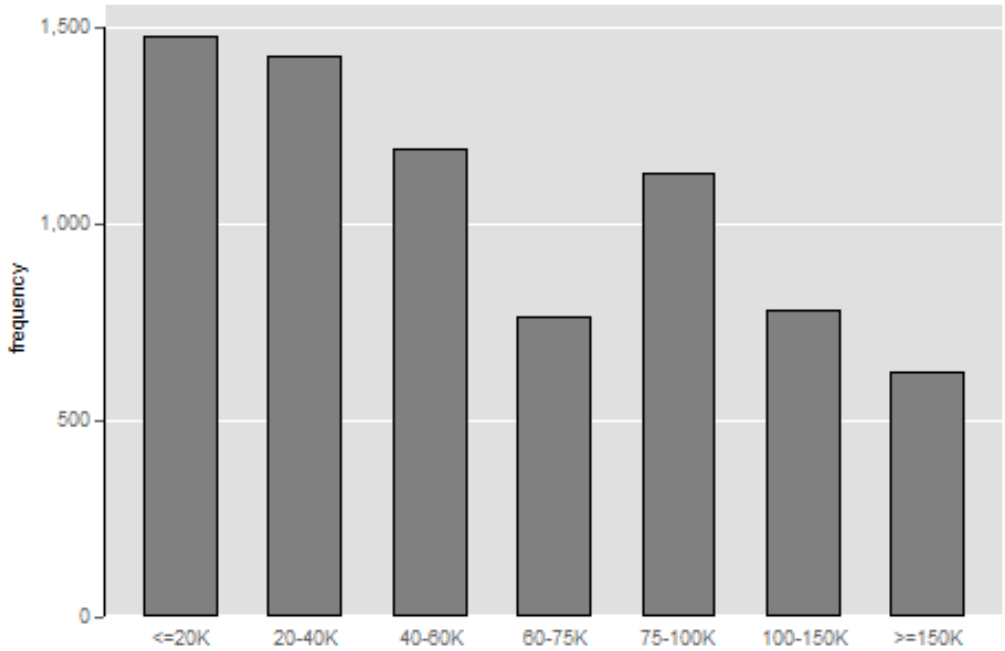
Pie Graphs



Code: `graph pie tucaseidr, over(hhincome)`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

- Requires a numeric variables that identifies observations.
- Takes counts of observations in each category of a categorical variable and presents them as proportions of all observations.

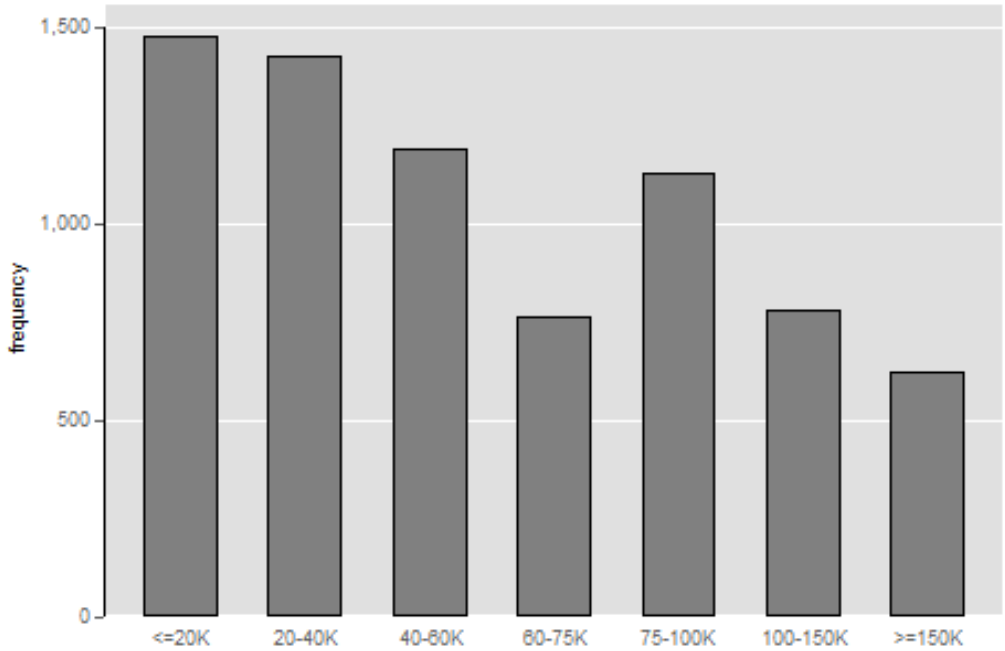
Bar graphs: Frequencies



Code: `graph bar (count), over(hhincome)`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

- Frequency graphs (using `(count)` in the code) only need a categorical variable defined. Conditions (using `if` statements) can be added before the comma.

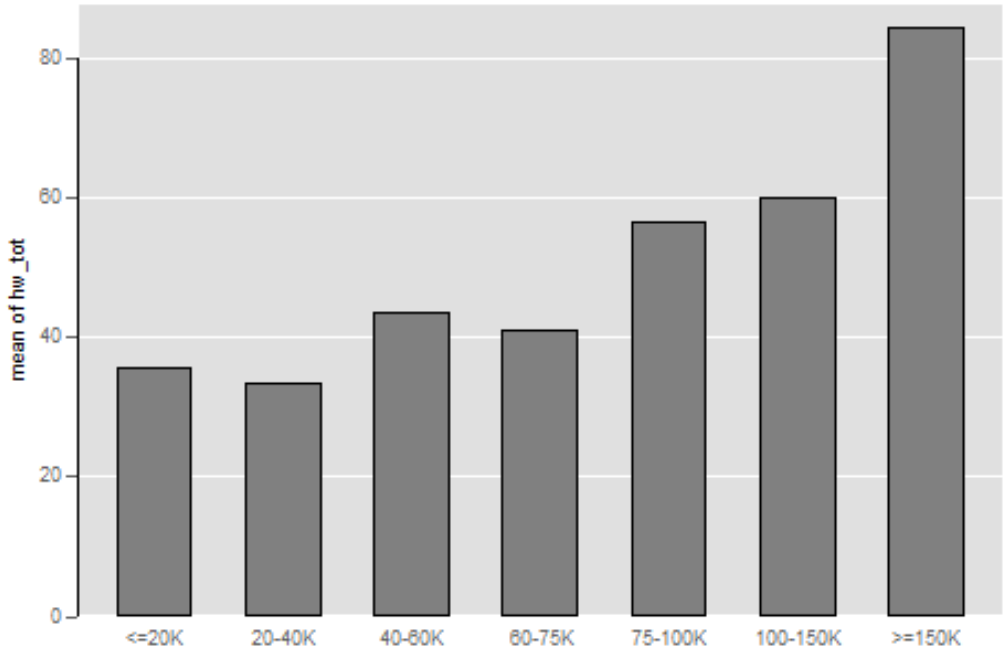
Bar graphs: Frequencies



Code: `graph bar (count), over(hhincome)`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

- Frequency graphs (using `(count)` in the code) only need a categorical variable defined. Conditions (using `if` statements) can be added before the comma.
- Takes counts of observations in each category of a categorical variable and presents them as bars.

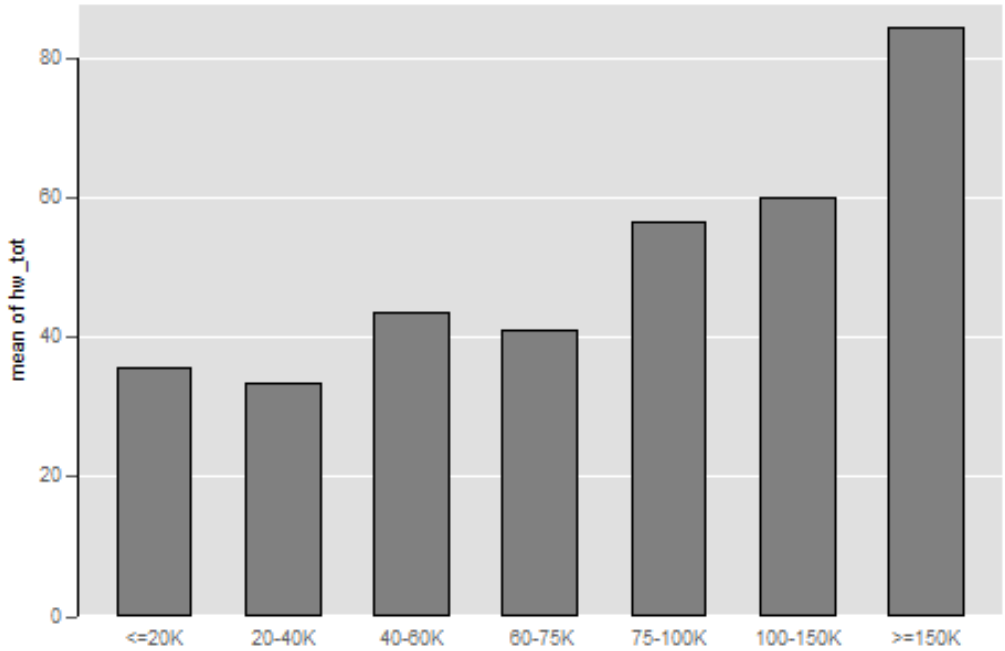
Bar graphs: Averages and Proportions



Code: `graph bar (mean) hw_tot, over(hhincome)`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

- Average graphs (using `(mean)` in the code) need a y-variable defined.

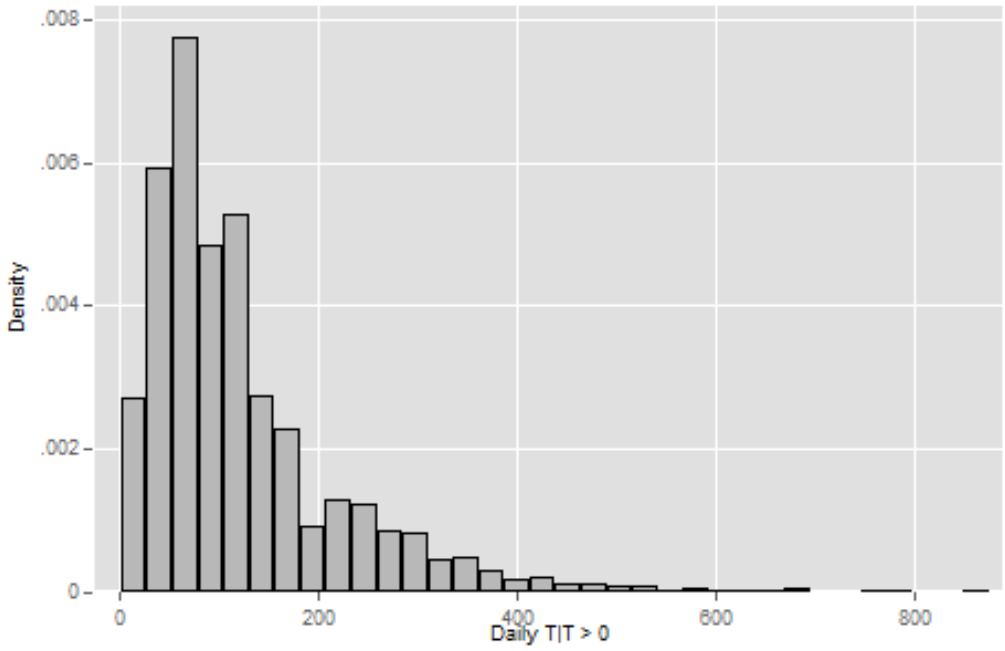
Bar graphs: Averages and Proportions



Code: `graph bar (mean) hw_tot, over(hhincome)`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

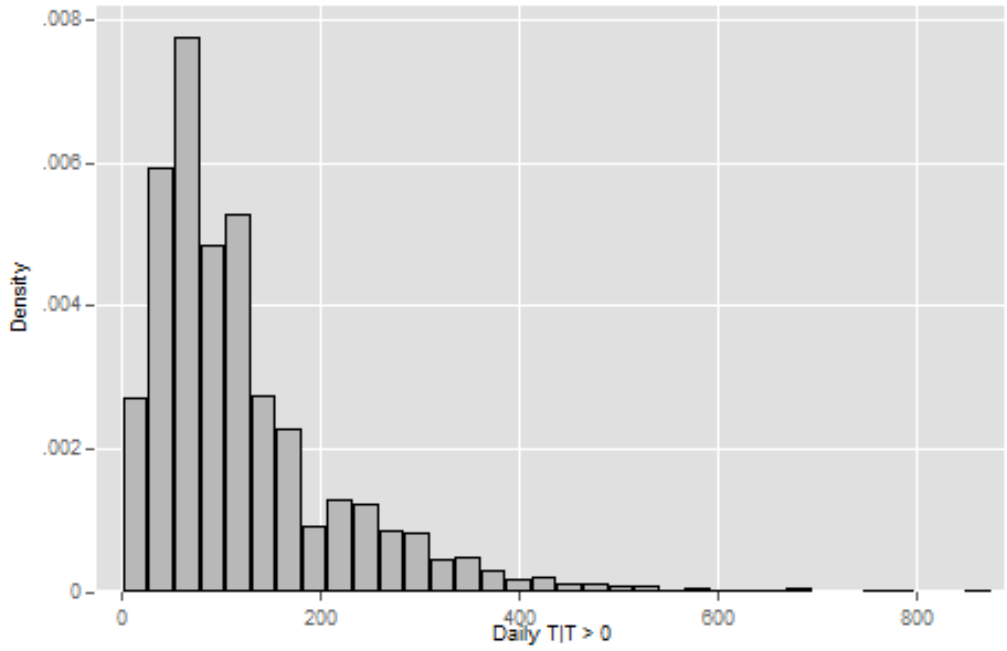
- Average graphs (using `(mean)` in the code) need a y-variable defined.
- Shows the average value of `yvar` (in this case `hw_tot`) within categories of a categorical variable. Provides proportions if `yvar` is indicator variable.

Histograms



Code: `histogram hw_tot2`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

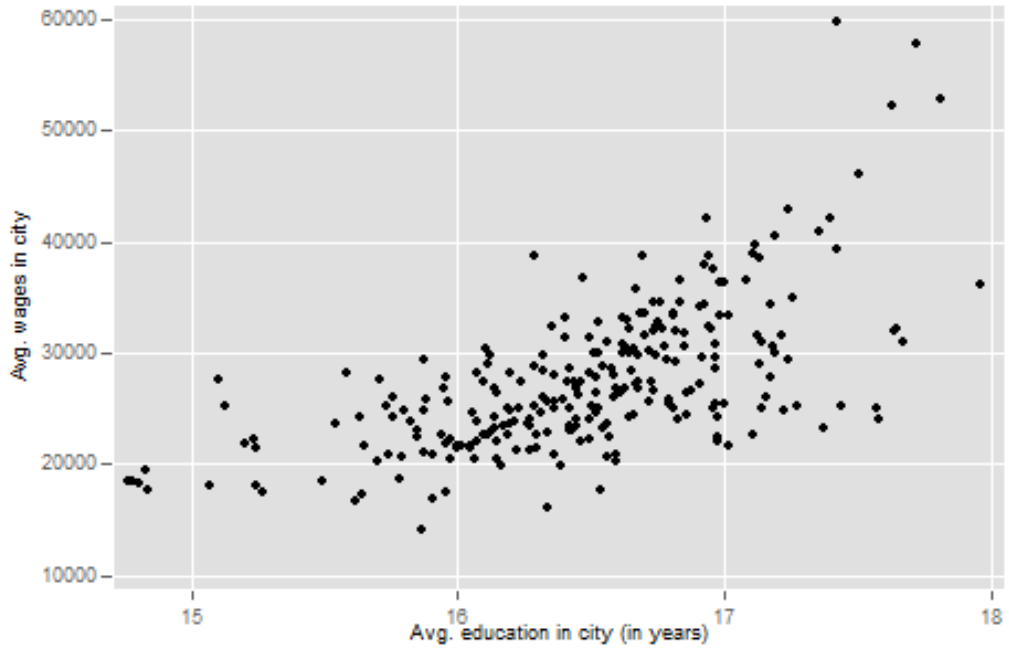
Histograms



Code: `histogram hw_tot2`
Source: ATUS data, Stata Lab in Week 1, Stata Handout 1

- Shows the distribution of values of `var` (in this canse `hw_tot2`) by presenting a count of observations in a given bin (i.e., range of values) for all possible values of `var`.

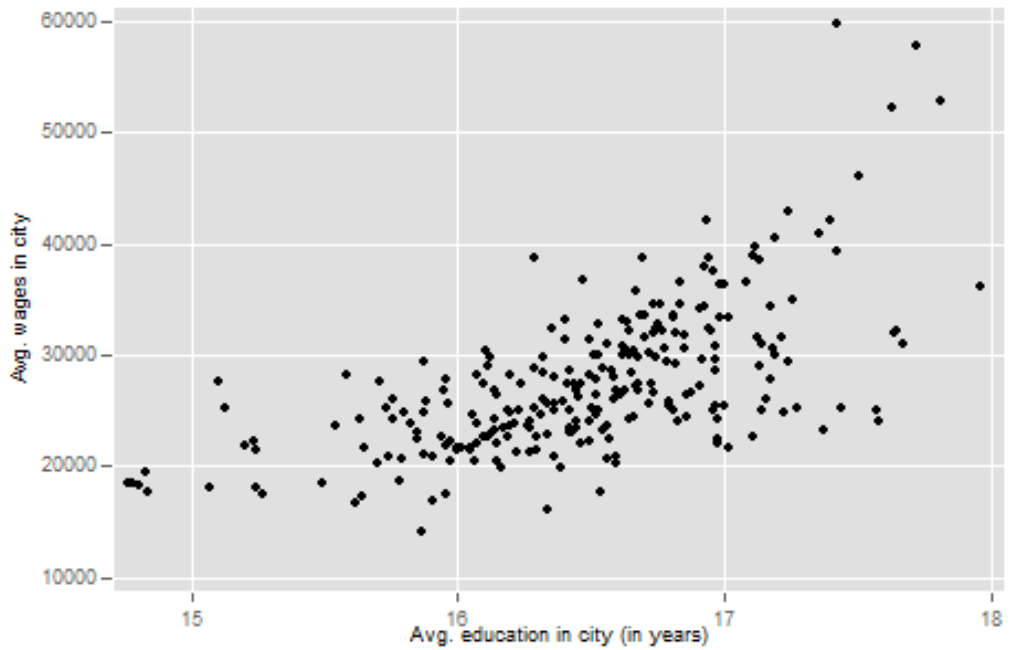
Scatterplots



Code: `scatter incwage_avg years_education_avg`
Source: ACS data - city level, Stata Lab in Week 2, Stata Handout 2

- Variable for the y-axis comes first in the code.

Scatterplots



Code: `scatter incwage_avg years_education_avg`
Source: ACS data - city level, Stata Lab in Week 2, Stata Handout 2

- Variable for the y-axis comes first in the code.
- Shows the value of `yvar` (here, `incwage_avg`) and `xvar` (here, `years_education_avg`) for each observation and plots each observation as a point on a coordinate plane. Useful for examining relationships between two variables.

Attendance