Dis 2: Descriptive Statistics; Sampling

Relevant textbook chapters: 4 and 5
Ch 4 and 5 handout and solution offered by Dr. Pac can be accessed here: Handout Solution
This handout incorporates reviews with all exercises from Dr. Pac's original handout.

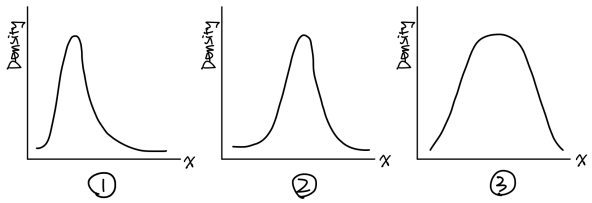
1 Motivation

- Last week, we talked about different types of data (interval vs. ordinal vs. nominal; or, population vs. sample), and what we can do with a set of data (descriptive statistics vs. inferential statistics).
- Like we mentioned last week, inferential statistics (using sample data to make inference about the population) is going to be our main focus for this course.
- In order to discuss inferential statistics, we first need to be able to understand our data, and be able to discuss how the sample data relate to the population data.
  - To understand the data (sample or population), we'll look at **descriptive statistics**.
  - To understand the link between sample and population data, we'll look at **sampling** techniques.

2 Descriptive Statistics

2.1 For a single variable (x)

- Recall that descriptive statistics are a set of methods that summarize or present your data.
- For example, say that you have three different sets of data distributed in the following way:



How can we tell the three data apart?

- **Method 1**: Use measures of central tendency

Name	**Population Notation**	**Sample Notation**	**Formula**
Mean	μ or μ_x	$\bar{x}$	Parameter: $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$ Statistic: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	-	-	Parameter and statistic: Sort all data, and take the middle one's value (or the average of the two middles)
Mode	-	-	Parameter and statistic: The most common observation(s)

- **Method 2**: Use measures of variation

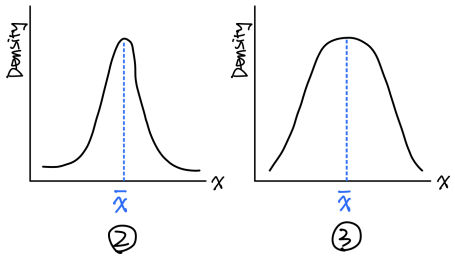
Name	**Population Notation**	**Sample Notation**	**Formula**
Variance	σ^2 or σ_x^2	s^2 or s_x^2	Parameter: $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$ $= \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu_x^2$ Statistic: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$
Standard deviation (std)	σ or σ_x	s or s_x	Parameter: $\sigma_x = \sqrt{\sigma_x^2}$ Statistic: $s_x = \sqrt{s_x^2}$

Side note: How does variance measure the variation of x?

Recall that for sample data, variance formula is

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Looking at the second and third data from earlier,



Clearly, data #3 has greater variation compared with data #2, so how is this reflected in the variance formula?

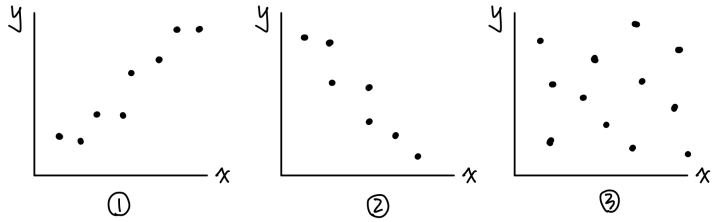
Side note: How does one interpret the standard deviation?

- **Empirical rule**: If the histogram of the data is **approximately bell-shaped**, then
  - Approximately 68% of all observations fall within one standard deviation of the mean
  - Approximately 95% fall within two standard deviations of the mean
  - Approximately 99.7% fall within three standard deviations of the mean
- **Chebysheff's Theorem** (more generally applied): the fraction of observations in **any** sample or population that lie within k standard deviations of the mean is at least

$$1 - \frac{1}{k^2} \quad \text{for} \quad k > 1$$

2.2 For two variables (x and y)

- Say that you now have three different sets of data with two variables, x and y, distributed in the following way:



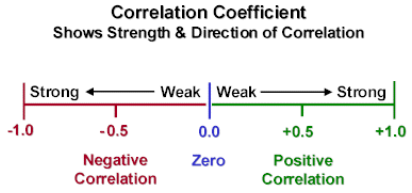
  How can we tell the three data apart?

- **Measures that describe the relationship between x and y:**

Name	**Population Notation**	**Sample Notation**	**Formula**
Covariance	σ_{xy}	s_{xy}	Parameter: $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$ $= \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right) - \mu_x \mu_y$ Statistic: $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ $= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}\right]$
Correlation / Correlation coefficient	ρ or ρ_{xy}	r_{xy} or $\hat{\rho}_{xy}$	Parameter: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ Statistic: $r_{xy} = \frac{s_{xy}}{s_x s_y}$

- What's the advantage of correlation coefficient compared with covariance?
  - Correlation coefficient rescales covariance by dividing covariance between x and y with the product of their standard deviations.
  - Therefore, correlation coefficient is always
    - Unitless
    - Between -1 and 1

    So we can compare correlation coefficient across datasets.
  - When correlation is **negative**, then x and y are **negatively correlated**.
    When correlation is **positive**, then x and y are **positively correlated**.
    When correlation is **exactly 0**, then x and y are **not correlated**.
    The closer the **absolute value** of correlation is to **1**, the **stronger** x and y's relationship is.
    The closer the **absolute value** of correlation is to **0**, the **weaker** x and y's relationship is.



(Image credit: The Significance of Correlation in Managed Futures)

3 Exercises: Descriptive Statistics

1. Consider the following data:

Student	**Grade (x)**	**Scoops of ice cream before the exam (y)**
1	1	0
2	2	0
3	2	1
4	3	1
5	3	1
6	3	0
7	4	0
8	4	1
9	5	1
10	6	2

(a) What is the variance of the students' grades (x) and scoops of ice cream (y)?

(b) In this particular sample, what fraction of the grade data (x) falls within 2 standard deviations of the mean?

Student	**Grade (x)**	**Scoops of ice cream before the exam (y)**
1	1	0
2	2	0
3	2	1
4	3	1
5	3	1
6	3	0
7	4	0
8	4	1
9	5	1
10	6	2

(c) Compare this actual result with the predictions of the Empirical Rule and with Chebysheff's Theorem.

(d) What is the interquartile range for the grade data (x)?

Student	**Grade (x)**	**Scoops of ice cream before the exam (y)**
1	1	0
2	2	0
3	2	1
4	3	1
5	3	1
6	3	0
7	4	0
8	4	1
9	5	1
10	6	2

(e) Calculate the covariance between grade (x) and ice cream consumption (y).

(f) Calculate and interpret the correlation between grade (x) and ice cream consumption (y).

2. Three professors are comparing grades of three classes for a midterm exam. Each class has 99 students.

   - In Class A: one student received grade of 1 point, another student got 99 points, and rest of the students scored 50 points.
   - In Class B: 49 students go a score of 1 point, one student got a score of 50 points, and 49 students got a score of 99 points.
   - In Class C: one student got a score of 1 point, one student got a score of 2 points, one student got a score of 3 points, one student got a score of 4 points, and so forth, all the way to 99.

   (a) Which class had the biggest average?

   (b) Which class had the biggest standard deviation?

   (c) Which class had the biggest range?

4 Sampling

- Recall from discussion 1 that a sample is a subset of data taken from the population. The action of taking this subset of data to construct your sample is called **sampling**.
- Eventually, our goal is to use the sample to draw conclusion about the population (inferential statistics), so it is important that our sample resembles the population.
- Different **sampling plans** are proposed to construct the sample, weighing the benefits of the plan against the costs:
  - **Simple random sampling**: every possible sample entry has equal chance of being selected.
  - **Stratified random sampling**: separate the population into mutually exclusive sets (i.e. strata), and then draw simple random samples from each stratum.
  - **Cluster sampling**: population is first divided into groups, and then one uses simple random sampling to select groups; all observations within the selected groups thus enter the sample.

Exercise. Which sampling plan is used in each of the following examples?

1. Categorize all Econ 310 students based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly selects 30 students from each class.

2. Categorize all Econ 310 students based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly select 2 out of the 5 possible groups. The groups corresponding with the class standing selected are chosen as the sample.

3. Number Econ 310 students sequentially from 1 to N. Draw 50 non-repeat random positive integers that are less than or equal to N. Select the students with the same numbers.

- As you can already see from the exercise, factoring in the specific steps taken when sampling, some sampling plan is expected to construct a sample that more closely resembles the population than the others.

  To formally examine how far the samples are from the population, we look at two types of errors that occur:

  1. **Sampling error**: difference between the sample and the population that exists only because the observations that happen to be included in the sample.
     $\Rightarrow$ increasing the sample size reduces this error
  2. **Nonsampling errors**: more serious type of error due to samples being selected improperly.
     $\Rightarrow$ increasing the sample size will NOT reduce this type of error

Nonsampling errors can be divided into three categories:

(a) **Errors in data acquisition**: the data is recorded wrong (due to incorrect measurement, mistake made during transcription, human errors)
(b) **Nonresponse errors**: responses are not obtained from certain people.
(c) **Selection bias**: some members from the target population cannot possibly be selected to be within the sample.

Exercise. Which type of error arises from the following examples?

1. You sent out a survey to all Econ 310 students via email, but some people quickly archived your email without filling out the survey.

2. You sent out a survey to all Econ 310 students via email, but some freshmen has yet to activate their UW email account, so the survey was not delivered to them.

3. You randomly selected 30 Econ 310 students to have them answer your survey questions. All 30 of them responded, and you did not make any mistake in recording the data. However, your result derived from the sample is still quite different from the parameter in the population.

4. You randomly selected 30 Econ 310 students to have them answer your survey questions. All 30 of them responded, but you messed up the order of items in two columns of the data recorded.