Lec 2*: Descriptive Statistics (Cont'd); Sampling

1 Motivation

- Last lecture, 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 class, inferential statistics (using sample data to make inference about the population) is going to be our main focus down the line.
- 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 (Cont'd)

2.1 For a single variable (x; cont'd from last time)

- Last lecture, we talked about two common sets of measures used to describe a single variable x:
  1. Use measures of central tendency (mean, median, mode)
  2. Use measures of variation (variance, standard deviation)
- Outside of central part of the data, and how widely the data points vary, another useful way to describe the data is to describe its **range**.
- **Method 3**: Use measures of range

Name	**Population Notation**	**Sample Notation**	**Formula**
Range	-	-	Parameter and statistic: Range $=$ Max $-$ Min
Interquartile range (IQR)	-	-	Parameter and statistic: IQR $=$ 75th percentile $-$ 25th percentile

- Side note: What is percentile? And how do you calculate the Pth percentile of x?
  - Think about something that you are likely already familiar with. When taking an exam (say, SAT), and the report card tells you that "your score is better than 95% of people", this means that your score is in the 95th percentile.
    - So, to loosely define percentile, say that there are 100 observations of x, then if you sort all observations from the smallest to the largest, the Pth percentile will correspond to the Pth observation of x, counting from the smallest observation to the largest.

*Some exercise questions are taken from or slightly modified based on Dr. Gregory Pac's Econ 310 discussion handout.

– For any set of data, the **location** of the Pth percentile value when data is sorted from smallest to largest (denoted as L_P) can be found by calculating

$$L_P = (n+1)\frac{P}{100}$$

* Once you have this location, you can then count to the L_P location, and the corresponding value is the Pth percentile.

Exercise. Define x variable as a variable that records people's favorite number. We have a sample of size 7. The sample data looks like the following:

$$10, 9, 50, 3, 16, 64, 1$$

What is the 25th percentile of this sample data?
Let's start off by sorting the sample data from the smallest to the largest:

$$1, 3, 9, 10, 16, 50, 64$$

Now, we need to find the location of the 25th percentile, so $P = 25$. Plugging into the location formula

$$L_{25} = (7+1) \times \frac{25}{100} = 8 \times \frac{1}{4} = 2$$

This means that the location of the 25th percentile, when data is sorted from the lowest, is the 2nd position. So the second data point when counting from the lowest is the 25th percentile of the sample data, meaning that the 25th percentile is 3.

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:

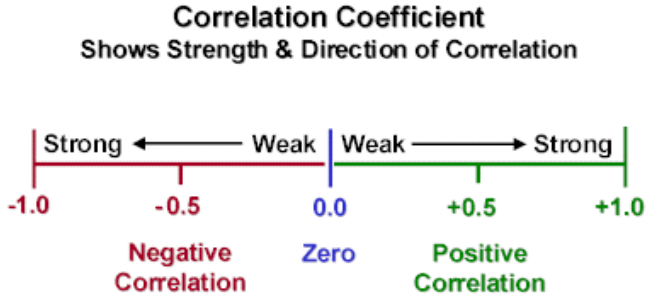
How can we tell the three data apart?

- **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)?

Using the simplified variance formula, one can calculate the variance for x and y as the following:

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{10-1}\left[129 - \frac{(33)^2}{10}\right] \approx 2.23$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{10-1}\left[9 - \frac{(7)^2}{10}\right] \approx 0.46$$

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

From (a), we found that the variance of x is 2.23. Taking the square root of variance gives us standard deviation. That is, the standard deviation of $x = s_x = \sqrt{s_x^2} = 1.49$.

The sample mean of x can also be calculated:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10} \times 33 = 3.3$$

Thus, when looking at "within 2 standard deviations of the mean" for x, we are looking at a range with a lower bound of

$$\bar{x} - 2s_x = 3.3 - 2 \times 1.49 = 0.32$$

and an upper bound of

$$\bar{x} + 2s_x = 3.3 + 2 \times 1.49 = 6.28$$

Going back to our data, every observation of x falls between 0.32 and 6.28. Thus, 100% of the

grade data fall within 2 standard deviations of the mean.

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

The interquartile range (IQR) is calculated in the following way:

$$IQR = \text{75th percentile} - \text{25th percentile}$$

The question now becomes how to find the relevant percentile measure. Luckily, the following formula has been suggested by your textbook (see Keller page 106):

$$L_P = (n+1)\frac{P}{100}$$

where P is the percentile number, and L_P denotes the location of the Pth percentile in a sorted data. For example, say that one wants to look up the location in the sorted data of the 10th percentile, then $P = 10$.

To find the 75th percentile, we first need its location in the sorted x data (notice that the x data presented in the table is already sorted, so we don't need to take a separate step to sort the data at the beginning). The location of the 75th percentile is

$$L_{75} = (10+1) \times \frac{75}{100} = 8.25$$

This tells us we should choose the value that's 0.25 of the distance between the eighth observation (4) and the ninth observation (5), so we get

$$\text{75th percentile} = 4 + (5-4) \times 0.25 = 4.25$$

Similarly, The location of the 25th percentile is

$$L_{25} = (10+1) \times \frac{25}{100} = 2.75$$

This tells us we should choose the value that's 0.75 of the distance between the second observation (2) and the third observation (2), so we get

$$\text{25th percentile} = 2 + (2-2) \times 0.75 = 2$$

Thus, $IQR = \text{75th percentile} - \text{25th percentile} = 4.25 - 2 = 2.25$

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

Using the simplified covariance formula, we have

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}\right] = \frac{1}{10-1}\left[29 - \frac{33 \times 7}{10}\right] \approx 0.66$$

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

From (b), we have that $s_x = \sqrt{s_x^2} = 1.49$

We can also calculate the standard deviation of y, which is $s_y = \sqrt{s_y^2} = 0.68$

This, along with the result from (e), allows us to calculate the correlation between x and y:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0.66}{1.49 \times 0.68} = 0.65$$

Since the correlation is positive, x and y have a positive relationship. Since its absolute value is relatively close to 1, x and y have a moderate level of relationship. Thus, we find a moderately strong positive correlation between grade (x) and ice cream consumption before the exam (y).

Keep in mind that the correlation coefficient alone does not tell us that this is a causal relationship (correlation is NOT causation).

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 has the biggest average?

   You can either calculate the mean of all three classes, or, given that all classes have the same number of students (99), you can simply look at whether all three classes have the same sum of grades among all 99 students (that is, whether $\sum_{i=1}^{99}$ is the same across all three classess).

$$\begin{aligned}
\sum_{i=1}^{99} \text{ in Class A} &= 1 + 99 + (99 - 2) \times 50 \\
&= 2 \times 50 + 97 \times 50 = 99 \times 50 \\
\sum_{i=1}^{99} \text{ in Class B} &= 49 \times 1 + 50 + 49 \times 99 = 49 \times (1 + 99) + 50 \\
&= 49 \times (2 \times 50) + 50 = (49 \times 2) \times 50 + 1 \times 50 = (49 \times 2 + 1) \times 50 = 99 \times 50 \\
\sum_{i=1}^{99} \text{ in Class C} &= 1 + 2 + 3 + \ldots + 97 + 98 + 99 = (1 + 99) + (2 + 98) + \ldots + (49 + 51) + 50 \\
&= 49 \times 100 + 50 = (49 \times 2) \times 50 + 50 = 99 \times 50
\end{aligned}$$

   Thus, all three classes have the same $\sum_{i=1}^{99}$, which means the mean $\frac{1}{99}\sum_{i=1}^{99}$ is the same across all three classes.

   (b) Which class has the biggest standard deviation?

   Based on the variance formula (since the standard deviation is just the squared root version of variance, we can look at the variance formula for some intuition), the standard deviation is large when there are many data points that are far away from the mean. In this case, Class B should have the highest standard deviation, as there are many more students far away from the average in this class.

(c) Which class has the biggest range?

Recall that range is calculated in the following way:

$$\text{range} = \text{highest observation} - \text{lowest observation}$$

Since the lowest score in all three classes is 1, and the highest score in all three classes is 99, range is the same across all three classes.

4 Sampling

- Recall from Lec 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 UW undergrads based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly selects 30 students from each class.

   Stratified random sampling
2. Categorize all UW undergrads 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.

   Cluster sampling
3. Number UW undergrads 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.

   Simple random sampling

- 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 UW students via email, but some people quickly archived your email without filling out the survey.

   Nonsampling error – nonresponse error

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

   Nonsampling error – selection bias

3. You randomly selected 30 UW 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.

   Sampling error

4. You randomly selected 30 UW 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.

   Nonsampling error – errors in data acquisition