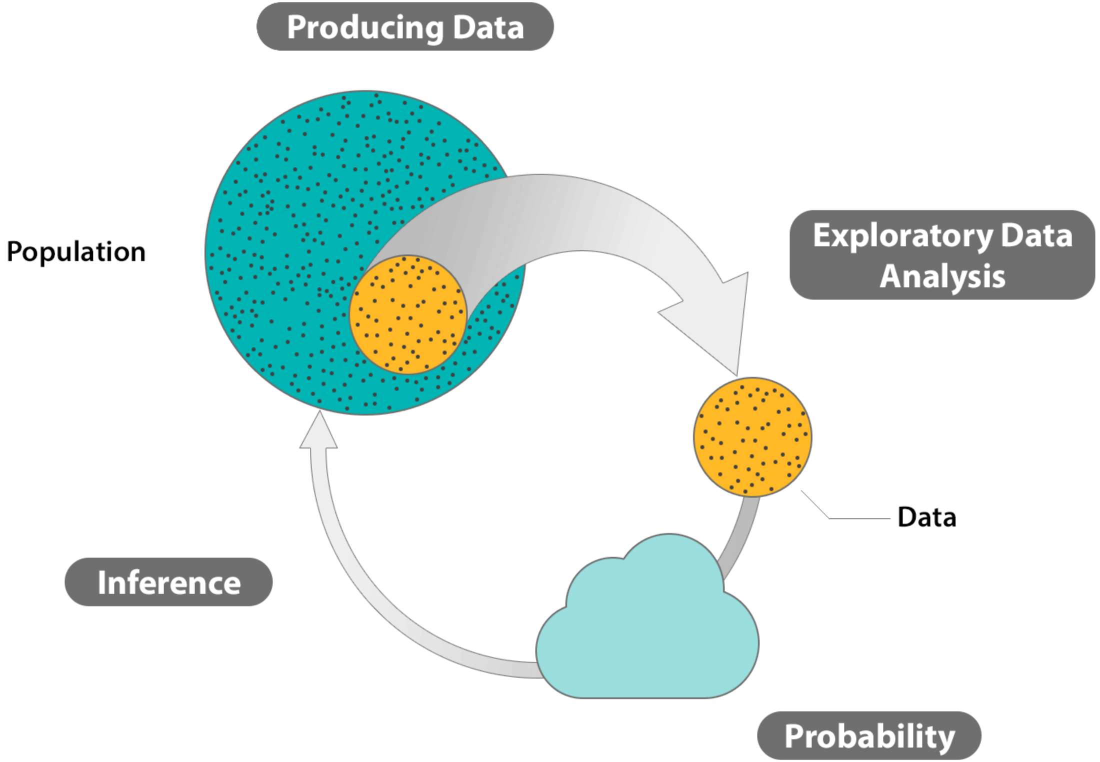
STATS 10: Final Review

Bingling Wang
Department of Statistics


Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability

Probability Wrap-Up

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Disjoint/Mutually exclusive events:

$$P(A \text{ and } B) = 0$$

$$P(A_1 \text{ or } A_2 \text{ or } \dots \text{or } A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Conditional probability: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

Independent events:

$$P(A \text{ and } B) = P(A)P(B)$$

$$P(A_1 \text{ and } A_2 \text{ and} \dots \text{and } A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n)$$

The Law of Large Numbers -- The reason why estimating theoretical probabilities from empirical probabilities works, and why simulations are useful.

- Implication
- Common misunderstanding

Practice

The game Monopoly is played by rolling a pair of dice. If you land in jail, then to get out, you must roll a double on any one of your next three turns, or else pay a fine. What are the chances that you get out of jail without paying a fine?

A multiple-choice exam has 10 questions. Each question has 3 possible answers, of which one is correct. A student knows the correct answers to 4 questions and guesses the answers to the other 6 questions. It turns out that the student answered the first question correctly. What are the chances that the student was merely guessing?

- $\dfrac{\left(\frac{1}{3}\right)\left(\frac{4}{10}\right)}{\left(\frac{1}{3}\right)\left(\frac{4}{10}\right)+(1)\left(\frac{6}{10}\right)}$
- $\dfrac{\left(\frac{1}{3}\right)\left(\frac{6}{10}\right)}{\left(\frac{1}{3}\right)\left(\frac{6}{10}\right)+(1)\left(\frac{4}{10}\right)}$
- $\dfrac{\left(\frac{1}{3}\right)\left(\frac{4}{10}\right)+(1)\left(\frac{6}{10}\right)}{\left(\frac{1}{3}\right)\left(\frac{4}{10}\right)}$
- $\dfrac{\left(\frac{1}{3}\right)\left(\frac{6}{10}\right)+(1)\left(\frac{4}{10}\right)}{\left(\frac{1}{3}\right)\left(\frac{6}{10}\right)}$

Probability Distribution Wrap-Up

1. Random variables:

- *Discrete random variable* - table, histogram, or a formula
  - The sum of the probabilities of all possible values must sum to 1.
- *Continuous random variable* - probability density curve
  - The probability of an interval is measured by the area under the curve
  - The total area under the curve must be 1.

2. Normal distribution

- Center, spread and shape
- Using a Z-table
  - Given a normal value x, solve for probability
  - Given a probability, solve for normal value x (percentile)

Sampling Distribution Wrap-Up

Statistics vary from sample to sample due to **sampling variability**, and therefore can be regarded as random variables whose distribution we call **sampling distribution**.

1. Behavior of sample estimates

   center, spread, shape

2. Accuracy and precision of the statistic
   - Bias
   - Standard error

3. Central Limit theorem

   Random and independent sample, Large sample size, Big population

			Sampling Distribution		
Variable	**Parameter**	**Statistic**	**Center**	**Spread**	**Shape**
Categorical (example: left-handed or not)	p = population proportion	$\hat{p}$ = sample proportion	p	$\sqrt{\frac{p(1-p)}{n}}$	Normal IF np ≥ 10 and n(1 - p) ≥ 10

Practice

Suppose that in Pet World, the population is 1000 people and 25% of the population are Cat People. Cat People love cats but hate dogs. We are planning a survey in which we take a random sample of 100 people. We calculate the proportion of people in our sample who are Cat People.

What value should we expect for this sample proportion? What's the standard error? How to interpret these values?

What is the approximate probability that the percentage in our sample will be bigger than 29%?

Confidence Interval Wrap-Up

Interval estimation: Estimate an unknown parameter using an *interval of values* that is likely to contain the true value of that parameter and state how confident we are that this interval indeed captures the true value of the parameter.

1. **CLT conditions** must be checked such that the sampling distribution of the statistic is approximately normal
2. **General form**

   $$\text{Point estimate} \pm \text{margin of error } (m)$$
   $$m = z^* \times SE$$

3. **Interpretation**

   The success rate of the method

4. **Effect of confidence level**

   Confidence level ↑, interval width ↑, precision ↓

5. **Determine sample size**

   To construct a 95% CI with a desired m, the sample size should be $\frac{1}{m^2}$

Practice

A recent study asked the question, “What percentage of children in California between the ages of 2 and 11 drink at least one sugary drink per day?” We wish to be able to calculate a 95% confidence interval for this estimate, and we want the margin of error to be 3 percentage points.

How many children should we randomly sample to achieve a margin of error of 3 percentage points if we use a 95% confidence level?

An actual study in California, based on a random sample of more than 1,111 children, estimated that a 95% confidence interval for the percentage of all children in California who drank at least one sugary drink per day was 28.3% to 33.5%, What is a margin of error in the interval constructed?

Practice

Do a majority of teachers in the U.S. think that digital devices (smartphones, tablets, computers) are mostly helpful for students' education? In 2018, Gallup took a poll of 497 randomly selected adults who teach K–12 students and 42% of them said that such devices had "mostly helpful" effects on students' education. (Source: Busteed & Dugan, 2018, news.gallup.com.)

Find a 95% confidence interval for the percentage of all K–12 teachers who believe that these devices have a mostly helpful effect on students' education. Is it plausible to conclude that as many as 50% or more teachers believe this?

Hypothesis Testing Wrap-Up

I. **Hypothesize**

H_0: no change, no difference (benefit of the doubt)
H_a: there is a change, there is a difference
Left-sided: “$<$”, Right-sided: “$>$”, Two-sided: “$\neq$”

II. **Summarize data**

Obtain data from sample and
1) Check the conditions
2) Calculate the test statistic z

III. **Assess evidence**

Left-sided test: p−value $= P(Z \leq z)$
Right-sided test, p−value $= P(Z \geq z)$
Two-sided test, p−value $= P(Z \leq -|z|) + P(Z \geq |z|)$

IV. **Conclude and interpret**

Given a certain significance level α,
1) Reject H_0, if p−value $\leq \alpha$
2) Do not reject H_0, if p−value $> \alpha$

Beware of the errors that could occur in hypothesis test!
Type I Error and Type II Error

Practice

Health professionals are often concerned about our lifestyles and how they affect our well-being. A group of medical researchers knew from previous studies that in the past, about 39% of all men between the ages of 45 and 59 were regularly active. Because regular activity is good for our health, researchers were concerned that this percentage had declined over time. For this reason, they selected a random sample, without replacement, of 1927 men in this age group and interviewed them. Out of this sample, 680 said that they were regularly active (Elwood. 2013).

Did the proportion of regularly active men decline?

	One Population Proportion
Parameter	p
Sample statistic	$\hat{p}$
Null hypothesis	$\mathrm{H_0}: p = p_0$ $\mathrm{H_a}$ three forms: $p < p_0$, or $p > p_0$, or $p \neq p_0$
CLT conditions	1. Random and independent 2. Large sample • **For Confidence Interval:** $\mathbf{n}\hat{\boldsymbol{p}} \geq \mathbf{10}, \mathbf{n}(\mathbf{1} - \hat{\boldsymbol{p}}) \geq \mathbf{10}$ • **For Hypothesis Test:** $\boldsymbol{np_0} \geq \mathbf{10}, \boldsymbol{n}(\mathbf{1} - \boldsymbol{p_0}) \geq \mathbf{10}$ 3. Large population $N \geq 10n$
Confidence interval	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Hypothesis Test	Z test statistic: $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ p-value: Left-sided: $\mathrm{P(Z \leq z)}$; Right-sided: $\mathrm{P(Z \geq z)}$; Two-sided: $\mathrm{P(Z \leq z) + P(Z \geq z)}$;