Announcement

- PS 1 uploaded is due Monday, January 27th.

Lecture 3

Statistics Review II

Outline

- Sampling distribution of the sample mean
- Central limit theorem
- Statistical inference-Confidence Interval (CI)
- Hypothesis testing

Population distribution vs. sampling distribution

- **Population distribution** of a random variable (X) is the distribution of its values for all members of the population.
  - Example: Height of individuals in the entire country.
- **Sampling distribution** is the probability distribution of a **statistic (e.g. mean ($\overline{X}$))**.
  - The average height of a class follows normal distribution-sampling distribution

Population and sampling distribution

	POPULATION	SAMPLING DISTRIBUTION
Mean	μ	$\mu_{\bar{x}} = \mu$
Standard Deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Shape	Normal	Normal
	Undetermined (skewed, etc.)	If n is "small" shape is similar to shape of original graph *OR* If n is "large" (rule of thumb: $n \geq 30$) shape is approximately normal (central limit theorem)

Statistical inference

- Statistical inference: draw conclusions about population from sample.
- Example
  - Parameter: average height of adults in the US
  - Statistic : average height of 50 randomly selected people from the population

- Suppose scores on an IQ test are normally distributed. 10 people are randomly selected and tested. The mean and standard deviation in the sample group is 95 and 15. Construct a 95% confidence interval for the true population mean.
- Extract information:
- n=10,degree of freedom=n-1=9; $\bar{X} = 95$;S=15; confidence level (C)=0.95

Steps to construct Confidence Interval(CI)

- Each confidence level C corresponds to a tail probability $\alpha = 1 - C$
- Given α and the degree of freedom (n-1), find $t_{\alpha/2}$ using the t-table
- Construct margin of error m=$t_{\alpha/2}$*s/$\sqrt{n}$
- CI=$\bar{X}$ $\pm$m=$\bar{X}$ $\pm$ $t_{\alpha/2}$*s/$\sqrt{n}$

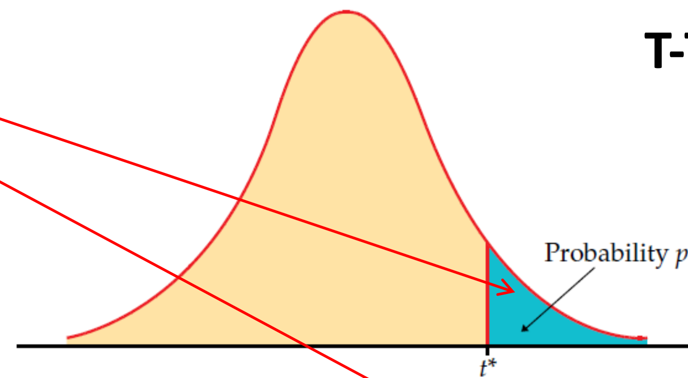
- Where s is the standard deviation of the sample, and $\bar{X}$ is the mean. n is sample size.

Answer

- Step 1: Convert C=0.95 to t-score using t table.
- C=0.95, $\alpha = 1 - C$=0.05, $\frac{\alpha}{2} = 0.025$
- look it up in t-table, $t_{\alpha/2}$=2.262
- Step 2: Construct margin of error.
- m=$t_{\alpha/2}$*s/$\sqrt{n}$=2.262*15/$\sqrt{10}$=10.73
- Step 3: Construct CI
- $\bar{X}$ $\pm$m=95 $\pm$10.73
- <u>**Translate the CI into real world meaning**</u>

- You are 95 percent confident that the true mean is within 84.27-105.73

T-Table



	Upper-tail probability p						
df	.25	.20	.15	.10	.05	.025	.02
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359

Test of significance (hypothesis test)

Test of significance (hypothesis test)

Example: The average hourly wage of a certain industry is said to be \$13. You randomly interviewed 49 firms in the industry and found the average hourly wage is \$12.5 and standard deviation is 2. So you start to question the number 13. Perform test of significance using α=0.10.

Extract Information: $\bar{X}$=12.5, s=2, n=49, df=48, $\mu = 13$, α=0.10

Steps to perform test of significance

- Step1: State null (H_0) and alternative hypotheses

  H0: $\mu=13$

  Ha: $\mu \neq 13$ ($\mu>13$ or $\mu<13$)

- Step2: Use test statistic to examine the compatibility of observed data with H_0

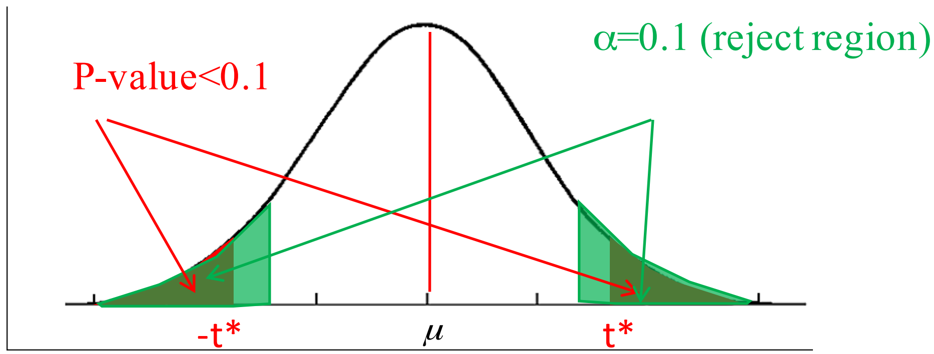
  $$t^*=\left|\frac{\bar{X}-\mu}{s/\sqrt{n}}\right|$$

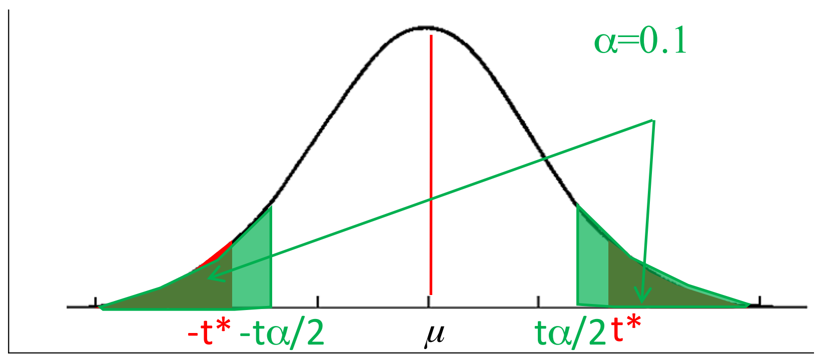
- $\bar{X}$=12.5, s=2, n=49, df=48, $\mu = 13$. Plug them into t, we get $t^*=\left|\frac{12.5-13}{2/\sqrt{49}}\right|=1.75$

Two ways of conducting hypothesis test

- P-value approach:
  - Translate t* into p-value(using T-table)
  - Compare P-value with α
- Critical value approach:
  - Translate α into $t_{\alpha/2}$(using T-table)
  - Compare t* with $t_{\alpha/2}$

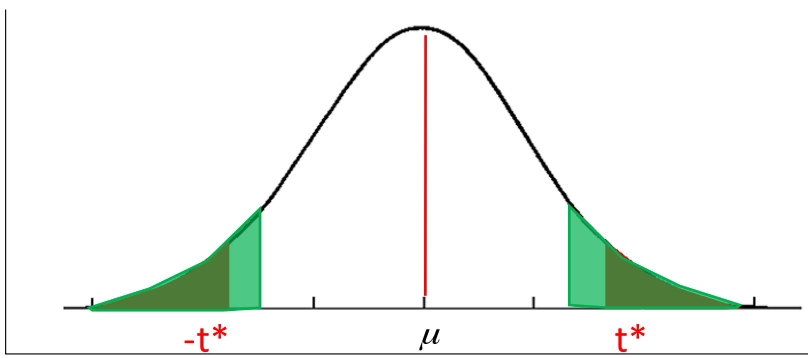
The bigger t* is, the smaller p-value is. The more likely you reject the null





Some tips to help you remember the rule

- To reject H_0, we want our sample to be far away from the middle-unusual sample based on H_0
- What does being far away mean?
  - a bigger t-value
  - a smaller p-value
- How far away is far enough?
  - Depend on α



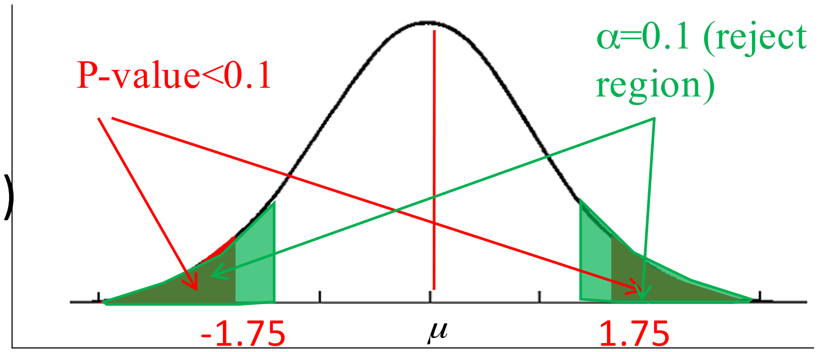
P-value approach

- Test statistic t*=| $\frac{\bar{X}-\mu}{s/\sqrt{n}}$|
- $\bar{X}$=12.5, s=2, n=49, df=48, $\mu = 13$. Plug them into t, we get t*=|$\frac{12.5-13}{2/\sqrt{49}}$|=1.75
- Step 3: Find p-value

Look t up in the [t-table,](t-table)

0.025<P(t>1.75)<0.05

P-value=2P(t>1.75) is within (0.05,0.1)



- Compare p with α

α=0.1, p-value<0.1, so α>p.

If p < alpha, we can reject H_0. The true average hourly wage is not 13.

Critical-value approach

- Compare t values instead of comparing probabilities

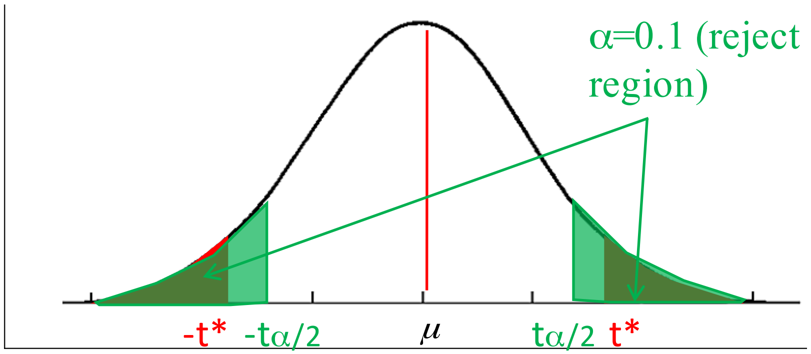
-Use the t-table to find the critical value of t at α=0.1, 1.684>$t_{\alpha/2}$>1.676

-compare t*=1.75 with t-critical, t*>$t_{\alpha/2}$

-Reject H_0

Takeaways:

The bigger t* is, the smaller p-value is. The more likely you reject the null



Logic of test of significance

- Start with a null hypothesis-e.g. Average height for female is 6 foot
- Based on the evidence-e.g. sample of heights of females in this class
- Conclude whether it is likely that the null is true-unlikely in this case

In the previous example:

- We assume average hourly wage is 13 (the null)
- Statistics tells us if wage is indeed 13, then the probability of seeing our current sample (p-value) is lower than 10% (α)
- Conclude-the null is unlikely to be true

Confidence interval and hypothesis test

- Confidence interval and hypothesis test are two sides of the same coin!
- Solve the question using confidence interval

The average hourly wage of a certain industry is said to be \$13. You randomly interviewed 49 firms in the industry and found the average hourly wage is \$12.5 with a standard deviation of 2. So you start to question the number 13.

Calculate confidence interval using C=0.9 and check whether 13 falls into the interval. Use CI to help you decide whether to reject H_0.

Steps to perform test of significance

- Step1: State null (H_0) and alternative hypotheses
- Step2: Use test statistic to examine the compatibility of observed data with H_0
- Step3: Two approaches
- a. P-value approach: Translate test statistic into P-value, and compare P-value with significance level α. State conclusion
- b. Critical-value approach: Find the t-critical value directed by α, and compare the test statistic with t-critical value. State conclusion

Steps to perform test of significance

- Step1: State null (H_0) and alternative hypotheses
- Step2: Use test statistic to examine the compatibility of observed data with H_0
- Step3: Two approaches
- a. P-value approach: Translate test statistic into P-value, and compare P-value with significance level α. State conclusion
- b. Critical-value approach: Find the t-critical value directed by α, and compare the test statistic with t-critical value. State conclusion

Confidence interval and hypothesis test

- Confidence interval and hypothesis test are two sides of the same coin!
- Solve the question using confidence interval

The average hourly wage of a certain industry is said to be \$13. You randomly interviewed 49 firms in the industry and found the average hourly wage is \$12.5 with a standard deviation of 2. So you start to question the number 13.

Calculate confidence interval using C=0.9 and check whether 13 falls into the interval. Use CI to help you decide whether to reject H_0.

Introduction to regression model (population)

TABLE 1.1 Donut Consumption and Weight

Observation number	Name	Donuts per week	Weight (pounds)
1	Homer	14	275
2	Marge	0	141
3	Lisa	0	70
4	Bart	5	75
5	Comic Book Guy	20	310
6	Mr. Burns	0.75	80
7	Smithers	0.25	160
8	Chief Wiggum	16	263
9	Principal Skinner	3	205
10	Rev. Lovejoy	2	185
11	Ned Flanders	0.8	170
12	Patty	5	155
13	Selma	4	145

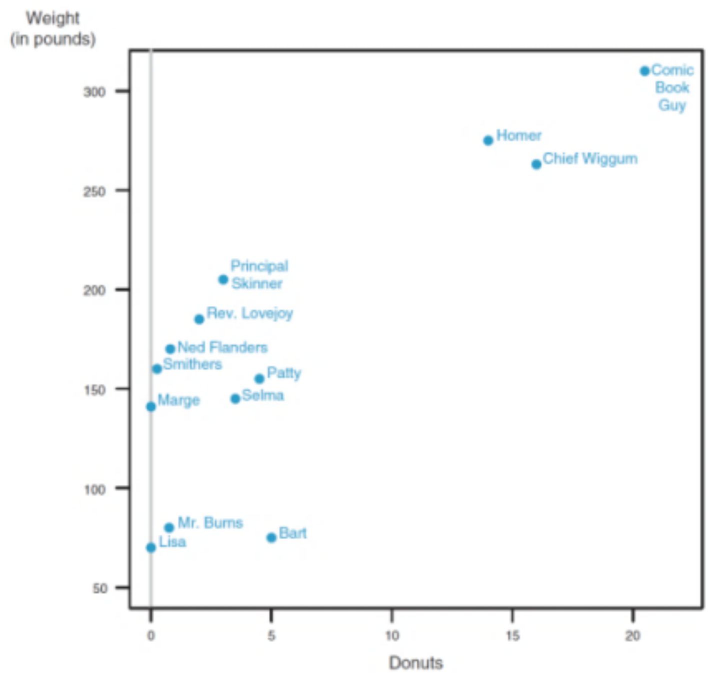


FIGURE 1.2: Weight and Donuts in Springfield

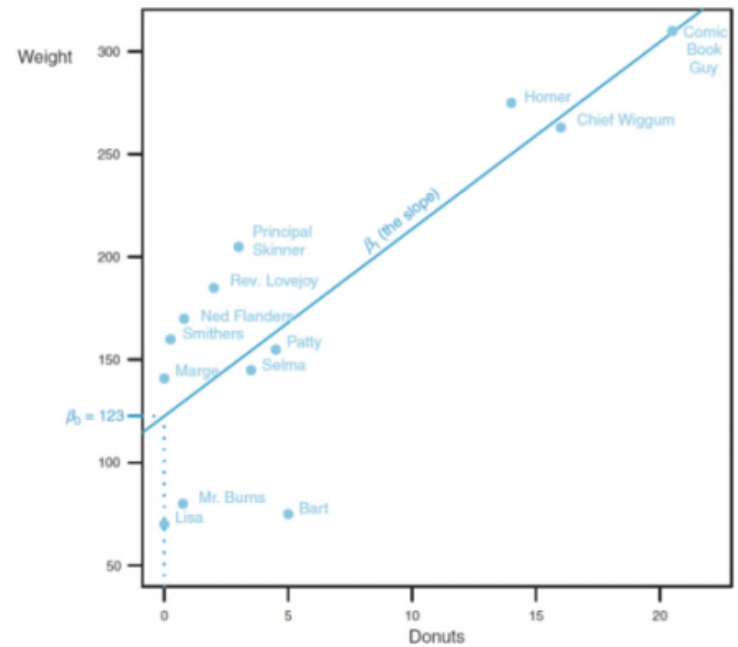
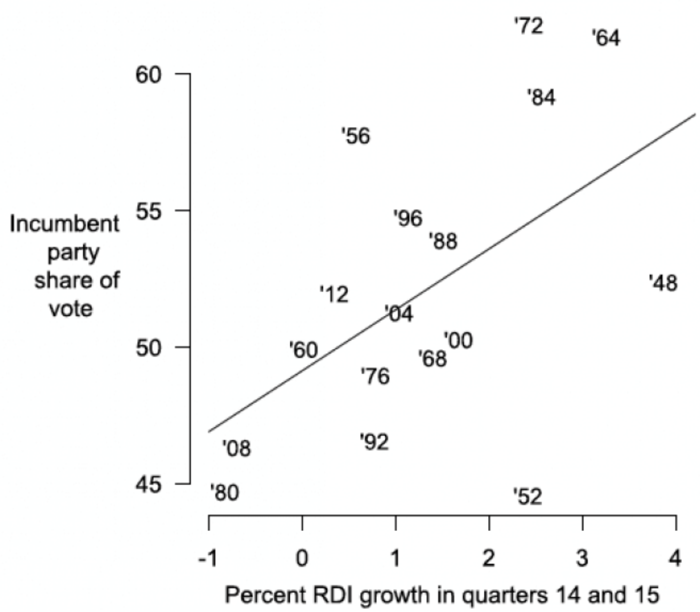


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

$$Weight_i = \beta_0 + \beta_1 Donuts_i + \epsilon_i \tag{1.1}$$

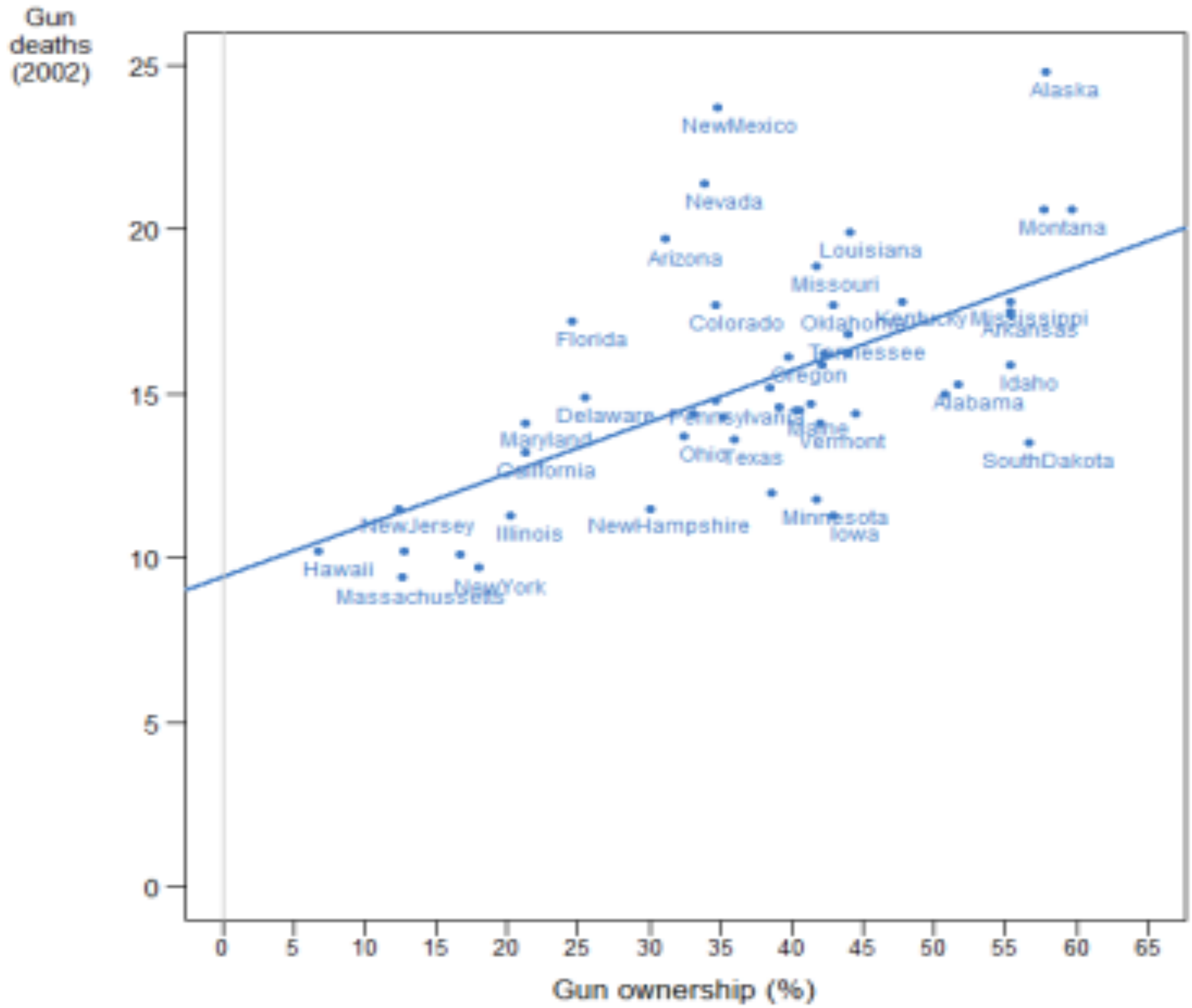
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.2}$$



$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.2}$$

Source: [MonkeyCage blog]

Gun deaths



Gun deaths (2002)
Gun ownership (%)
Alaska
NewMexico
Nevada
Montana
Louisiana
Arizona
Missouri
Colorado
Florida
Tennessee
Oregon
Idaho
Alabama
Delaware
Maryland
California
Ohio
Texas
Vermont
SouthDakota
NewJersey
Illinois
NewHampshire
Iowa
Hawaii
Massachusetts
NewYork
0
5
10
15
20
25
30
35
40
45
50
55
60
65

CASE STUDY

Flu Shots

$$Death_i = \beta_0 + \beta_1 Flu\ shot_i + \epsilon_i \tag{1.3}$$

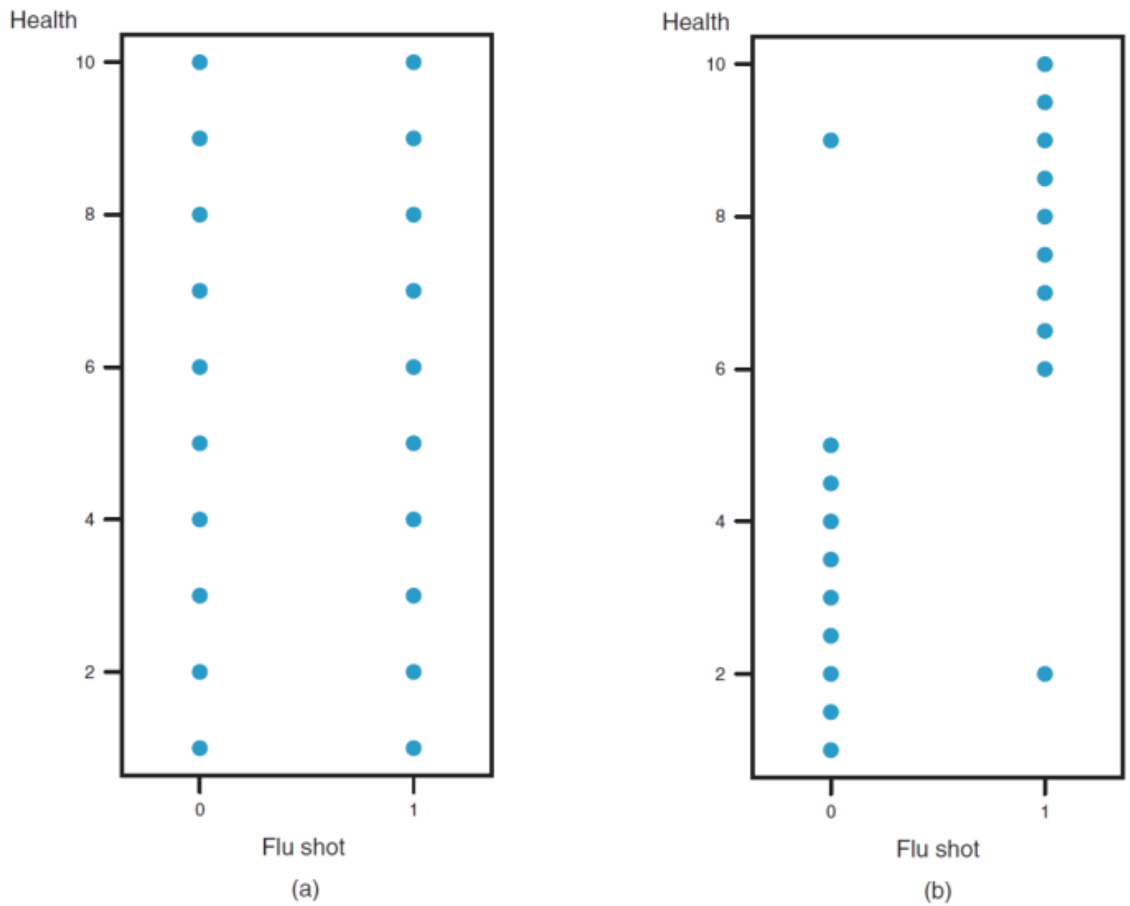
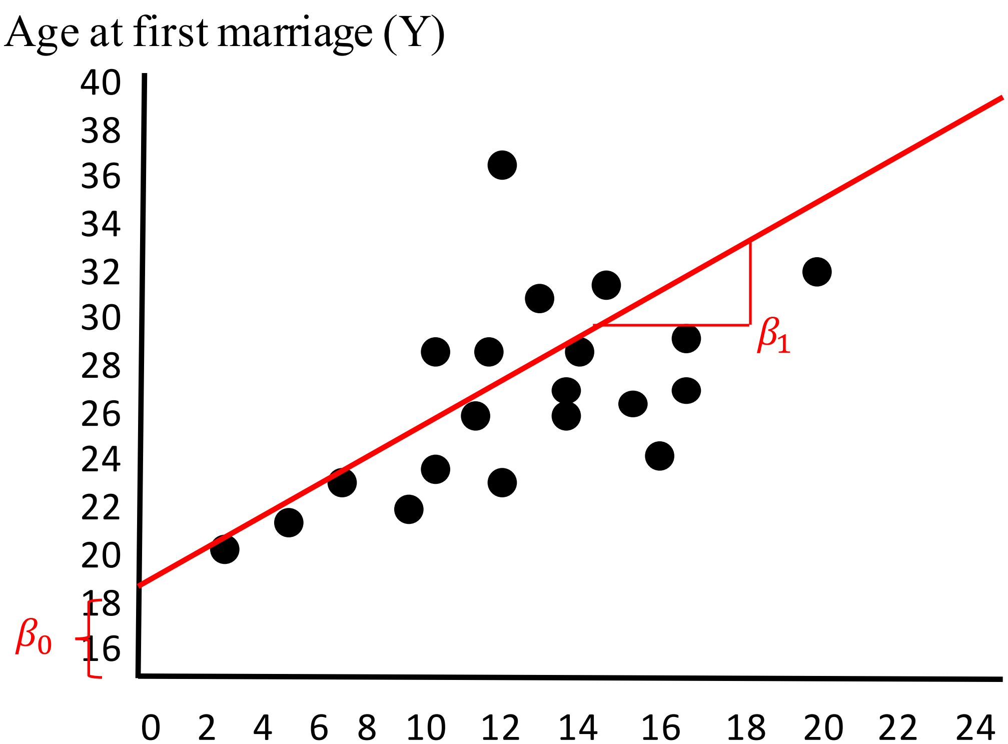


FIGURE 1.7: Two Scenarios for the Relationship between Flu Shots and Health

Research question: will more education defer marriage?

- What are some possible channels that education might affect marriage?
- What are the two key variables in this example?

Visually display the relationship



$Y = \beta_0 + \beta_1 X$

This is NOT the regression model, but close!

$Slope = \beta_1 = \frac{\Delta Y}{\Delta X}$

REMEMBER THIS

Our core statistical model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

1. β_1, the slope, indicates how much change in Y (the dependent variable) is expected if X (the independent variable) increases by one unit.
2. β_0, the intercept, indicates where the regression line crosses the Y-axis. It is the value of Y when X is zero.
3. β_1 is almost always more interesting than β_0.

Review for quizzes

- Know how to use CI and hypothesis test to perform statistical inference
- Understand the concepts of statistical significance, reject/fail to reject the null hypothesis
- Know the regression model in the population