Sample Statistics vs. Population Parameters

Sample Statistics:

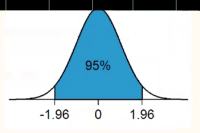
- Functions of your sample:
  - Only gives information about the sample, may not provide the entire story about the population.
- Notation:
  - $\bar{X}$, S^2 (S), $\hat{p}$

Population Parameters:

- Describes the entire population:
  - Usually given to you, there is no clear cut way to calculate a population parameter from a given sample.
- Notation:
  - μ, $\sigma^2(\sigma)$, p

给一个值算interval

Confidence Intervals



- Margin of Error
  - How to calculate:
    - (Critical Value)*(Standard Error)
- Confidence Interval:
  - How to calculate
    - Point Estimate $\pm$ (Critical Value)*(Standard Error)
    - For the sample mean: $\bar{X} \pm (Z^*)(\sigma / \sqrt{n})$
- What it means:
  - For a 95% confidence interval, we are 95% confident that the **true mean** (i.e. population mean) lies within this interval.
- How to interpret values of interest compared to your confidence interval? (With regards to hypothesis testing)***
  - Inside confidence interval → You **would not** be surprised if the value of interest occurred.
  - Outside confidence interval → You **would** be surprised if the value of interest occurred.

***You can't make any concrete conclusions about the significance of the value (or whether or not to reject the null hypothesis) from the confidence interval.

Critical Value (Z*)
95% - qnorm(1 - (1 - 0.95) / 2) ~ 1.96
99% - qnorm(1 - (1 - 0.99) / 2) ~ 2.58

whether or not be surprised if true mean

Mean or Proportion for Parameter of Interest

Mean is for numeric values

Proportion is for categorical values

Examples: Would you use mean or proportion? why?

> Studying if students who attend the review session have a higher letter grade than those who don't. catagorical

> Looking at the number of hours of sleep students get by class level. numerical

分辨案例

Hypothesis Testing

Normal Hypothesis Testing

- Test Statistic:
  - Understand the context of the question and determine what your test statistic is going to be (mean or proportion?).
- Compare your test statistic to your Normal distribution. $P(Z \geq z_{observed}) =$ p-value
  - P-value: the probability that a random variable value would occur by chance.
- Compare p-value to significance level
  - ($\alpha = 0.05$)
  - What happens when your p-value is below 0.05? reject null hypothesis

T-Test

- Observations are claimed to have come from a normal distribution.
  - (use when standard error or standard deviation is unreliable) ie small sample size
- Test Statistic:
  - Determine if you are dealing with sample mean, sample proportion, difference of means, etc.
  - Compare your test statistic to the T-Distribution. Remember that Degrees of Freedom = n - 1
  - Compare p-value to significance level (α)
    - ($\alpha = 0.05$)

Hypotheses

Null H_0:

Usually states the status quo, that there is no difference or that the data matches the given value.

Ex: **H_0: observed value = population mean**

Alternative H_a:

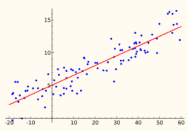
Our research question, a statement that contradicts the null hypothesis.

Ex: **H_a: observed value ! = population mean**

Regression

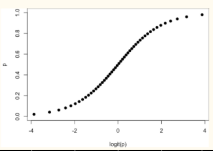
Linear

- Response variable is **numerical**.
- Model: $Y = \beta_0 + (\beta_1 * X_1) + ... + (\beta_n * X_n)$
  - $X_1, ..., X_n$ correspond to your predictor variables and β_0 is the intercept of the line.



Logistic

- Response variable is **categorical**.
- High p-hat → Success likely; low p-hat → failure likely. prediction
- Model: $\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...)}}$



Different types of Hypothesis Testing

Single mean: We are focused on only one dataset and on how this data's mean corresponds to the population mean. E.g. does the average SAT score at our school match the national average?

Difference of two means: Focuses on the difference between two observed sample means and tries to answer the question whether the two population means are different.

Difference of paired data: Data is ***paired*** when each observation in one data set has a special correspondence with exactly one observation in the other data set.
Ex. Is there a difference between the price of a book at UCLA and the price of the *same* book on Amazon?

Single Proportion: Does the proportion of our sample differ from the known proportion? E.g. The known number of defectives is 10%, if 100 randomly selected items from a particular batch produce 20 defectives, is this batch's defective rate significantly different.

Difference of two proportions: Is there a difference between two population proportions, given two sample proportions. E.g. Do Factory A and Factory B produce a similar amount of defective items?

连线题 给案例选types

R^2

R^2 tells us the proportion of the variability in the response variable that is explained by the explanatory variable.

To find R^2, you just square R, the correlation coefficient

Say the R^2 value is 0.679. We would interpret it like this:

"67.9% of the variability in <response variable, y> is explained by <explanatory variable, x>."

- Note that from the R-output, the 'Multiple R-squared' or 'Adjusted R-squared' can be used.

Which hypothesis test is required?

- Does the average SAT score at our school match the national average? Single mean
- Is there a difference between the price of a book at UCLA and the price of the *same* book on Amazon? Diff of paired data
- The known number of defectives is 10%, if 100 randomly selected items from a particular batch produce 20 defectives, is this batch's defective rate significantly different. Single proportion
- Is student attendance higher at UA basketball games than UCLA basketball games? Diff of 2 means
- Do Factory A and Factory B produce a similar amount of defective items? Diff of 2 proportions

好难道写null hypo的题

Writing H_0 and H_A in Symbols

- **Difference of Two Means**
  - (H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 \neq 0$)
- **Paired Data**
  - (H_0: $\mu_{diff} = 0$ H_a: $\mu_{diff} \neq 0$)
- **Single Proportion**
  - (H_0: $p = p_0$ H_a: $p \neq p_0$) known proportion
- **Difference of Two Proportions**
  - (H_0: $p_1 - p_2 = 0$ H_a: $p_1 - p_2 \neq 0$)
- **Linear Regression**
  - (H_0: $beta_x = 0$ H_a: $beta_x \neq 0$) (x - "insert variable here") slope <no trend?>

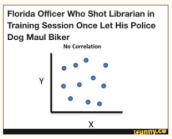
Conditions for Inference

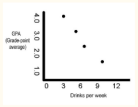
	Normal	T-Test	Regression	Single Proportion	Difference of Proportions
Independence	• Random sampling • Independent variables • Less than 10% of population	• Random sampling • Independent variables • Less than 10% of population	• Residuals with constant variance • Residuals distribution nearly normal	• Random sampling • Independent variables • Less than 10% of pop.	• Random sampling • Independent variables • Less than 10% of pop.
Size/Shape	• At least 30	• Often used for samples of lower than 30	• Linear association between each explanatory variable (Xs) and the response variable (Y)	• 10 successes and failures within sample distribution	• 10 successes and failures w/ pooled proportion
Other	• Not a lot of skew		• Linear for numeric, logistic for categorical		

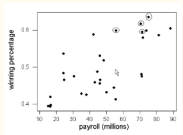
出案例
① 判断类型
② 看 condition met 与否
直接指着一个一个看

Correlation

Visual Understanding:







Numerical Understanding:

- Correlation Coefficient (R):
  - Strength of association between response and explanatory variables
  - [-1, -0.5) → Strong to moderate negative Correlation
  - (-0.5, 0.5) → Weak Negative or Positive Correlation
  - (0.5, +1] → Moderate to Strong positive Correlation

Pooled Proportions

A company is trying to determine if a new production method produces fewer defects than their current method.

	Defective	Functional
New Method	5	40
Old Method	25	30

What is a "success"?

Calculate the pooled estimate. Are conditions met for hypothesis testing for normal distribution?

Practice - Determining the Test Statistic (4.1)

- In a survey, one hundred college students are asked how many hours per week they spend on the Internet mean – numeric
- In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your coursework?" mean – percentages
- In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers. proportion Y/N categorical
- In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within 1 year after graduation. same above

Paired or not? (5.17)

In each of these examples, determine if the data are paired.

a. Compare pre- (beginning of semester) and post-test (end of semester) scores of students. same student 2 data paired
b. Assess gender-related salary gap by comparing salaries of randomly sampled men and women. not paired
c. Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E of the same group of patients. paired
d. Assess effectiveness of a diet regimen by comparing the before and after weights of subjects. paired

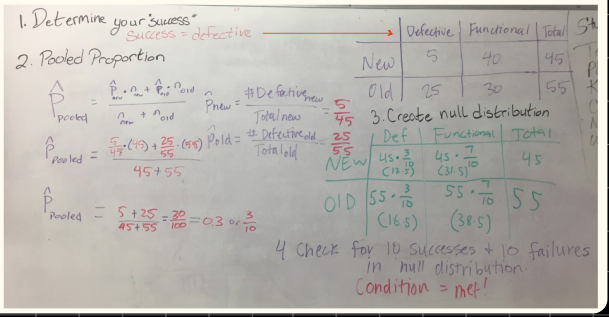
判断 proportion

Practice - Hypothesis Testing

Determine what kind of hypothesis test would be used for each and write out the null and alternative hypotheses for the following studies. (First in words then in symbols)

- A random sample of 25 New Yorkers were asked how much sleep they get per night. Do these data provide convincing evidence that New Yorkers on average sleep less than 8 hours a night? Single mean ~~yes/no~~ sleep & hrs / not
- We believe we will see an increase in test scores from the mid-term to the final. Looking at all the test scores in the classes, do they provide convincing evidence that there is an increase in scores from the mid-term to the final? ~~2 means~~ paired ~~yes/no~~ H0: diff = 0
- A company is trying to determine if a new production method produces fewer defects than their current method. Diff of 2 prop no/yes proportions same
- Do older people have a higher salary? ~~diff 2 means~~ yes/no linear regression beta age = 0

Answers to previous problem



$\hat{p}_{new} = \frac{5}{45}$ $\hat{p}_{old} = \frac{25}{55}$ $\hat{p}_{pooled} = \frac{\hat{p}_{new} * n_{new} + \hat{p}_{old} * n_{old}}{n_{new} + n_{old}}$ 45 55

abcd 原题 直接看答案

Baby Weights pt III (8.3)

a. Write out the equation for the model
b. What percent of the variance of the response variable is explained by the model?
c. Explain how we should interpret the coefficients for **gestation** and **age** in this context.
d. What is the null hypothesis for interpreting the p-value for the **weight** (mother's) variable? Are we able to reject the null hypothesis?
e. **Bonus**: Which variables are statistically significant?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

Multiple R-squared: 0.2504

This linear regression is for predicting the average birth weight based on different factors.

Birth weight (hat) = -80.41 + 0.44 * gestation + -3.33 * parity + ... + -8.40 * smoke

25.04% of the variance in birth weight is explained by the model.

0.44 oz increase in birth weight for every additional day of pregnancy, when all else is held constant.

-0.01 oz decrease in the birth weight for every additional year in the mother's age.

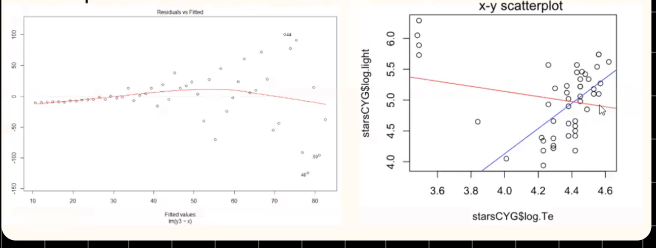
D. betaweight = 0

Yes, we are able to reject it because the p-value for weight is 0.0471 and that is less than alpha of 0.05

E. gestation, parity, height, weight, smoke

4道选择，看p值 → 0.8 → 0

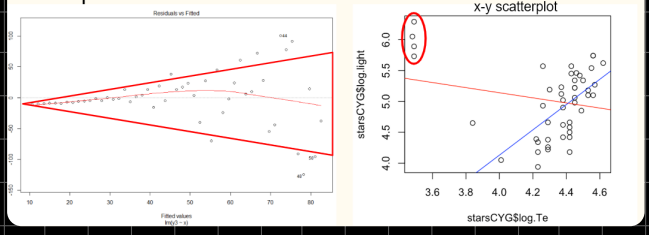
What's going on with these graphs? Would they be a problem for the linear model?



① not consistent variance
"cone shape"

② high influence, high leverage 4 points

What's going on with these graphs? Would they be a problem for the linear model?



Chronic Illness Pt. II (4.9)

Continued from last problem... Identify each of the following statements as True or False. **Justify** each of your answers.

a. We can say with certainty that the confidence interval from ex. 4.7 contains the true percentage of US adults who suffer from chronic illness.
b. If we repeated this study 1000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of US adults who suffer from chronic illnesses.
c. The poll provides statistically significant evidence (at the $\alpha = 0.05$ level) that the percentage of US adults who suffer from chronic illnesses is below 50%.
d. Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

Practice - Confidence Intervals (4.7)

Chronic Illness Pt. I In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions". However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study

```
> conf_interval <- 0.45 + c(-1.96*0.012, 1.96*0.012)
> conf_interval
[1] 0.42648 0.47352
```

(a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time.

(b) True. Notice that the description focuses on the true population value.

(c) True. If we examine the 95% confidence interval computed in Exercise 4.9, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5.

(d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.

问：logical regression & linear regression 的区别
google 一下，发现确实的