STA 371G: Statistics and Modeling

Review of Basic Probability and Statistics: Estimation and Sampling Distributions

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/STA371G

Sampling Distribution of Sample Mean

- To estimate the average starting salary of BBA graduates
- We assume that the starting salary follows a certain distribution, with μ as the population mean and σ^2 as the population variance.
- It is impossible/expensive to ask all BBA graduates to estimate the true values of μ and σ^2
- We take a random sample of $n = 100$ BBA graduates $x_1, \cdots, x_n$, based on which we estimate μ with the sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- How accurate is $\bar{X}$ as an estimate of μ?

Sampling Distribution of Sample Mean

- The expectation of $\bar{X}$:

$$E(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n} E(X_i) = \mu$$

- The variance of $\bar{X}$:

$$Var(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n^2} Var(X_i) = \frac{\sigma^2}{n}$$

Let $X_1, \cdots, X_n$ be normal random variables that $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ also follows a normal distribution:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

This is called the sampling distribution of the sample mean...

Sampling Distribution of Sample Mean

- The sampling distribution of the sample mean $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ describes how our estimate of the population mean μ would vary over different samples of the same size n
- It provides us with a vehicle to evaluate the uncertainty associated with our estimate of the mean...
- In practice, the population variance σ^2 is also usually unknown.
- How to estimate σ^2?

Sampling Distribution of Sample Mean

- The sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- It turns out that s^2 is a good proxy for σ^2, so we can approximate the sampling distribution by

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{s^2}{n}\right)$$

- We call $\sqrt{\frac{s^2}{n}}$ the standard error of $\bar{X}$... it is a measure of its variability... I like the notation

$$s_{\bar{X}} = \sqrt{\frac{s^2}{n}}$$

Sampling Distribution of Sample Mean

Approximately, we have

$$\bar{X} \sim \mathcal{N}\left(\mu, s^2_{\bar{X}}\right)$$

- $\bar{X}$ is unbiased... $E(\bar{X}) = \mu$. On average, $\bar{X}$ is right!
- $\bar{X}$ is consistent... as n grows, $s^2_{\bar{X}} \to 0$, i.e., with more information, eventually $\bar{X}$ correctly estimates μ!

Central Limit Theorem (Optional)

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ follows a normal distribution:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- What if X does not follow a normal distribution?
- Let $X_1, \cdots, X_n$ be a random sample from a distribution (any distribution!) with a finite mean μ and a finite variance σ^2, the Central Limit Theorem tells us that as the sample size n increases, the distribution of the sample mean $\bar{X}$ approaches the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
- We will demonstrate this using Simulation with R.

Sampling Distribution of Sample Variance (Optional)

If $X_1, \cdots, X_n$ is a random sample from a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ and $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ is the sample variance, then

- summation of squared iid standard normal random variables:

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$$

- the sampling distribution of the sample variance:

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

- s^2 is an unbiased estimate of the population variance σ^2:

$$E\left(\frac{(n-1)s^2}{\sigma^2}\right) = n - 1 \quad \Rightarrow \quad E(s^2) = \sigma^2$$

Student t Distribution (Optional)

Recall that $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

- If we know the population variance, then we estimate the confidence intervals using

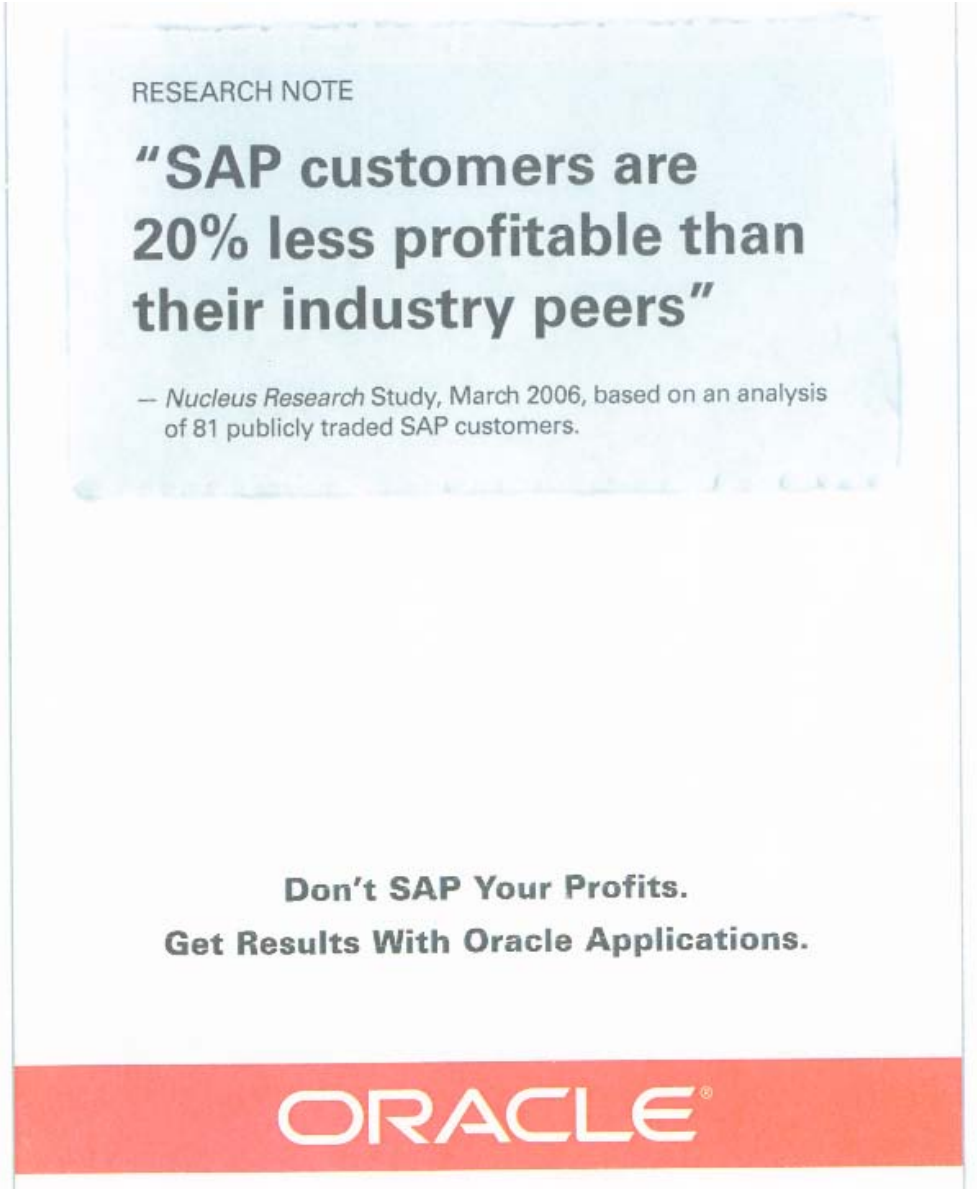
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

- If we do not know the population variance, then we estimate the confidence intervals using

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}} \sim t_{(n-1)}$$

As n increases, $t_{(n-1)}$ (t distribution with $n-1$ degrees of freedom) approaches the standard normal distribution. In this class, we assume that n is large enough so that we can safely use the standard normal distribution to measure uncertainties.

Oracle vs SAP Example (understanding variation)



Oracle vs. SAP

- Do we "buy" the claim from this add?
- We have a dataset of 81 firms that use SAP...
- The industry ROE is 15% (also an estimate but let's assume it is true)
- We assume that the random variable X represents ROE of SAP firms and can be described by

$$X \sim N(\mu, \sigma^2)$$

	$\bar{X}$	s^2
SAP firms	0.1263	0.065

- Well, $\frac{0.12}{0.15} \approx 0.8$! I guess the add is correct, right?
- Not so fast...

Oracle vs. SAP

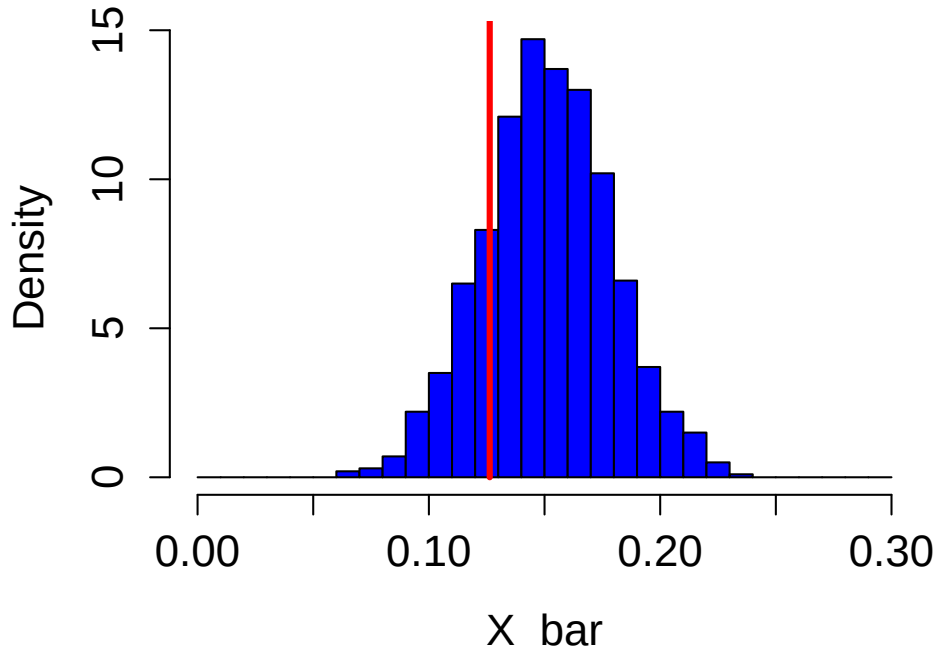
- Let's assume that the ROE of firms using SAP is, on average, the same as the industry. Assume further that s^2 is a good estimate of the variance...

$$ROE \sim N(0.15, 0.065)$$

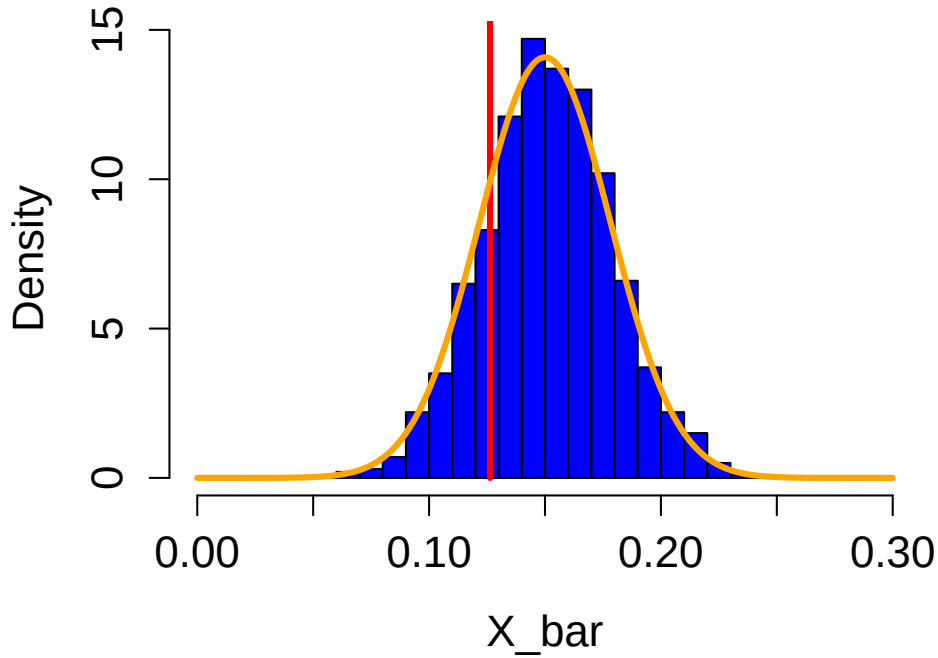
- In a random sample of 81 firms, how often can we expect the sample mean to be equal or below 0.1263?
  Using R to do the calculation:
  pnorm(0.1263,0.15,sqrt(0.065/81)) = 0.20
  pnorm((0.1263-0.15)/sqrt(0.065/81)) = 0.20
- What does this mean if we try to compare the profitability of firms using SAP versus the industry? Do we have strong evidence to support the claim of the study?

Oracle vs. SAP

- Let's do a little simulation...
- Generate 1000 different samples of size 81 from a $N(0.15, 0.065)$. Plot the histogram (density) of $\bar{X}$... Now, what do you think about the ad?



Our simulation was done assuming that $\mu = 0.15$... in that case $\bar{X} \sim N\left(0.15, \frac{0.065}{81}\right)$



Confidence Intervals

Approximately, we have

$$\bar{X} \sim N\left(\mu, s_{\bar{X}}^2\right)$$

so...

$$\frac{\bar{X} - \mu}{s_{\bar{X}}} \sim N(0, 1)$$

right?

- What is a good prediction for μ? What is our best guess??
  $\bar{X}$
- How do we make mistakes? How far from μ can we be??
  95% of the time $\pm 2 \times s_{\bar{X}}$
- $[\bar{X} \pm 2 \times s_{\bar{X}}]$ gives a 95% range of plausible values for μ... this is called the 95% Confidence Interval for μ.

Oracle vs. SAP example... one more time

In this example, $\bar{X} = 0.1263$, $s^2 = 0.065$ and $n = 81$... therefore, $s^2_{\bar{X}} = \frac{0.065}{81}$ so, the 95% confidence interval for the ROE of SAP firms is

$$\left[\bar{X} - 2 \times s_{\bar{X}}; \bar{X} + 2 \times s_{\bar{X}}\right]$$

$$= \left[0.1263 - 2 \times \sqrt{\frac{0.065}{81}}; 0.1263 + 2 \times \sqrt{\frac{0.065}{81}}\right]$$

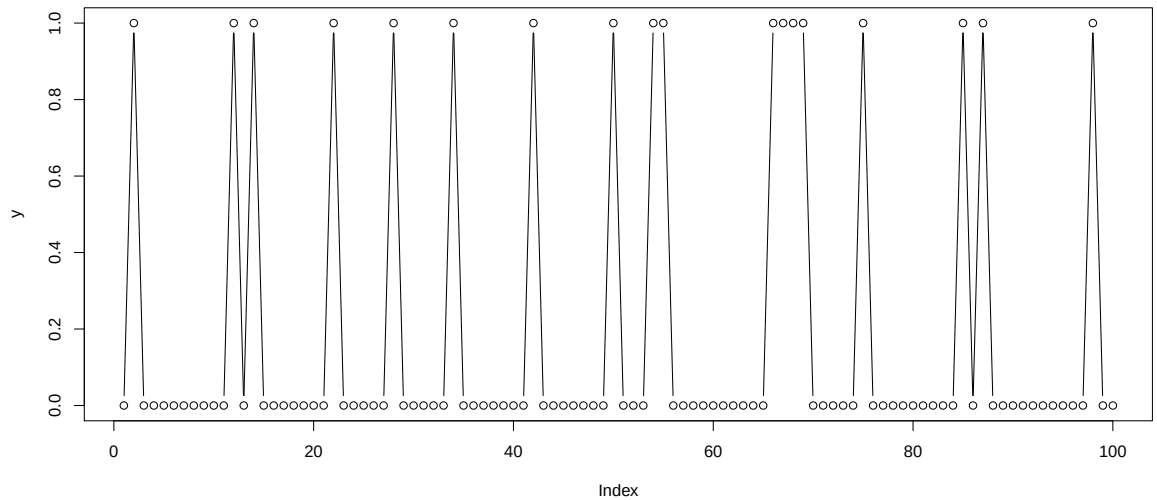
$$= [0.069; 0.183]$$

- Is 0.15 a plausible value? What does that mean?

Estimating Proportions... another modeling example

Your job is to manufacture a part. Each time you make a part, it is defective or not. Below we have the results from 100 parts you just made. $Y_i = 1$ means a defect, 0 a good one.

How would you predict the next one?



There are 18 ones and 82 zeros.

In this case, it might be reasonable to model the defects as iid...

We can't be sure this is right, but, the data looks like the kind of thing we would get if we had iid draws with that p!!!

If we believe our model, what is the chance that the next 10 are good?

$.82^{10} = 0.137.$

Sampling Distribution of Sample Proportion

- Let $X \sim \text{Binomial}(n, p)$ and define the sample proportion as

$$\hat{p} = \frac{X}{n}.$$

- When n is relatively large and p is not too close to 0 or 1:
  - X is approximately distributed as

$$X \sim \mathcal{N}(np, np(1-p))$$

  - The sample proportion $\hat{p}$ is approximately distributed as

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

We used the proportion of defects in our sample to estimate p, the true, long-run, proportion of defects.

Could this estimate be wrong?!!

Let $\hat{p}$ denote the sample proportion.

The standard error associated with the sample proportion as an estimate of the true proportion is:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

We estimate the true p by the observed sample proportion of 1's, $\hat{p}$.

The (approximate) 95% confidence interval for the true proportion is:

$$\hat{p} \pm 2\, s_{\hat{p}}.$$

Defects:

In our defect example we had $\hat{p} = .18$ and $n = 100$.

This gives

$$s_{\hat{p}} = \sqrt{\frac{(.18)\,(.82)}{100}} = .04.$$

The confidence interval is $.18 \pm .08 = (0.1, 0.26)$

Polls: yet another example...

If we take a relatively small random sample from a large population and ask each respondent "Vote for Democrat" or "Vote for Republican", where p is the true population proportion of "Vote for Democrat".

Suppose, as is common, $n = 1000$, and $\hat{p} \approx .5$.

Then,

$$s_{\hat{p}} = \sqrt{\frac{(.5)(.5)}{1000}} = .0158.$$

The standard error is .0158 so that the $\pm$ is .0316, or about $\pm$ 3%.

(Sounds familiar?!)

The Bottom Line...

- Estimates are based on random samples and therefore random (uncertain) themselves
- We need to account for this uncertainty!

- "Standard Error" measures the uncertainty of an estimate
- We define the "95% Confidence Interval" as

$$\text{estimate} \pm 2 \times \text{s.e.}$$

- This provides us with a plausible range for the quantity we are trying to estimate.

The Bottom Line...

- When estimating a mean the 95% C.I. is

$$\bar{X} \pm 2 \times s_{\bar{X}}$$

- When estimating a proportion the 95% C.I. is

$$\hat{p} \pm 2 \times s_{\hat{p}}$$