Section 2: Estimation and Confidence Intervals

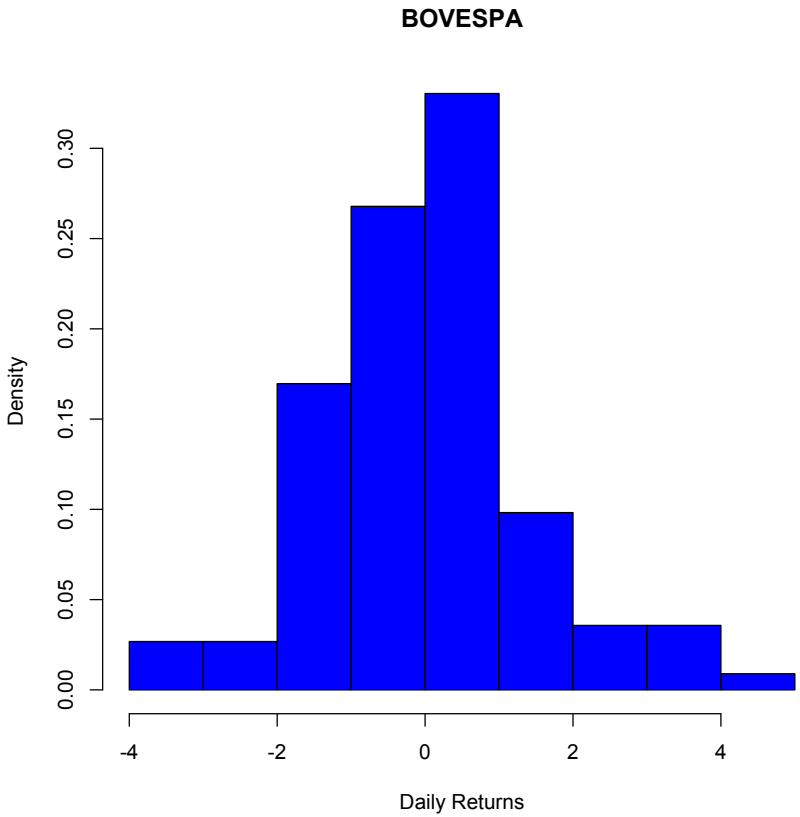
Carlos M. Carvalho
The University of Texas at Austin
McCombs School of Business

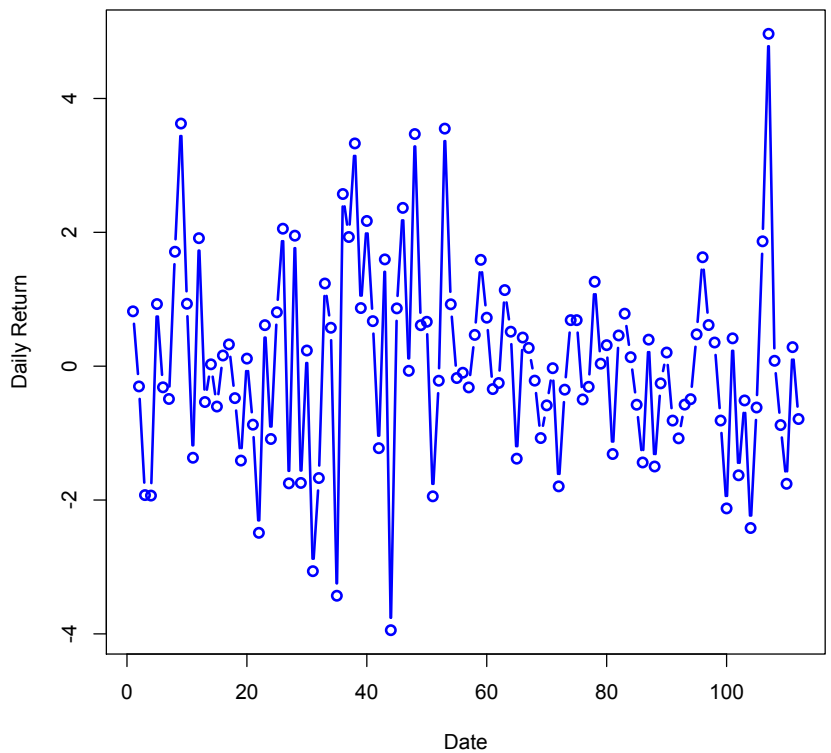
http://faculty.mccombs.utexas.edu/carlos.carvalho/teaching/

A First Modeling Exercise

- I have US$ 1,000 invested in the Brazilian stock index, the IBOVESPA. I need to predict tomorrow's value of my portfolio.
- I also want to know how risky my portfolio is, in particular, I want to know how likely am I to lose more than 3% of my money by the end of tomorrow's trading session.
- What should I do?

IBOVESPA - Data





- As a first modeling decision, let's call the random variable associated with daily returns on the IBOVESPA X and assume that returns are independent and identically distributed as

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

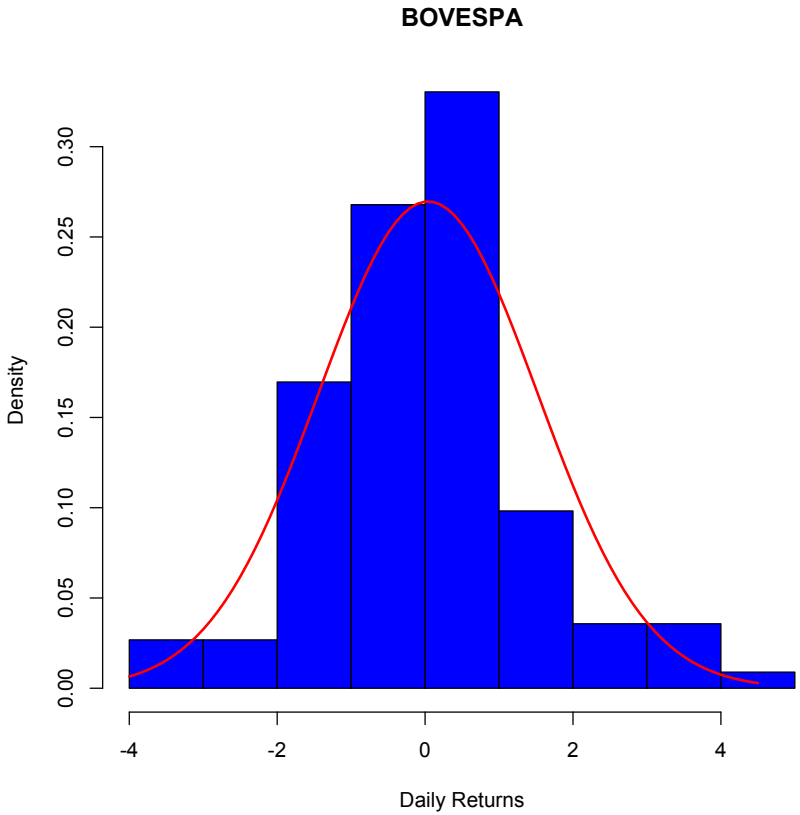
- Question: What are the values of μ and σ^2 ?
- We need to estimate these values from the sample in hands (n=113 observations)...

- Let's assume that each observation in the random sample $\{x_1, x_2, x_3, \ldots, x_n\}$ is independent and distributed according to the model above, i.e., $x_i \sim N(\mu, \sigma^2)$
- An usual strategy is to estimate μ and σ^2, the mean and the variance of the distribution, via the sample mean ($\bar{X}$) and the sample variance (s^2)... (their sample counterparts)

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

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

For the IBOVESPA data in hands,



$\bar{X} = 0.04$ and $s^2 = 2.19$

- The red line represents our "model", i.e., the normal distribution with mean and variance given by the estimated quantities $\bar{X}$ and s^2.
- What is $Pr(X < -3)$?

Annual Returns on the US market...

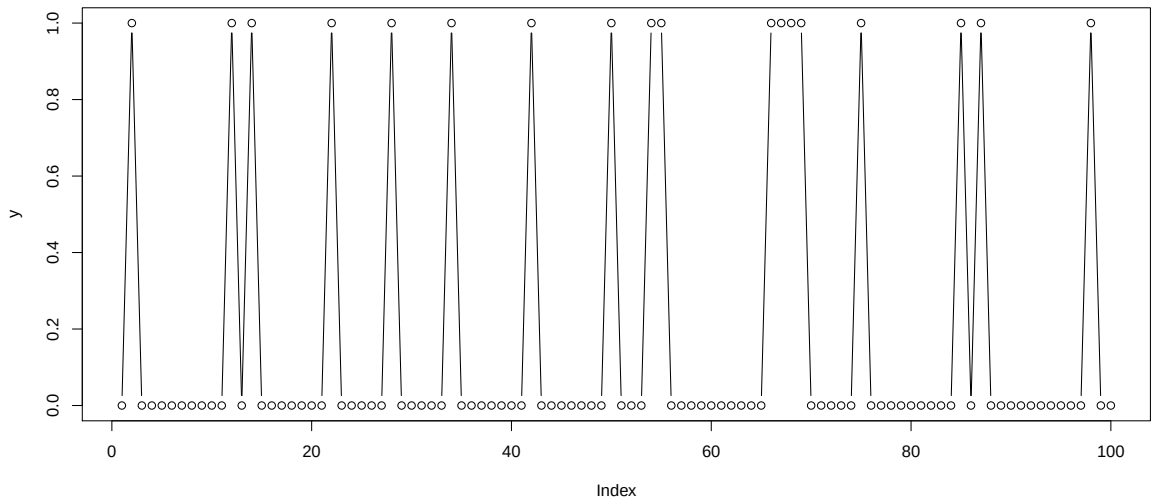
Assume I invest some money in the U.S. stock market. Your job is to tell me the following:

- what is my expected one year return?
- what is the standard deviation (volatility)?
- what is the probability my investment grow by 10%?
- What happens in 20 years if I invest \$1 today on the market?

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?

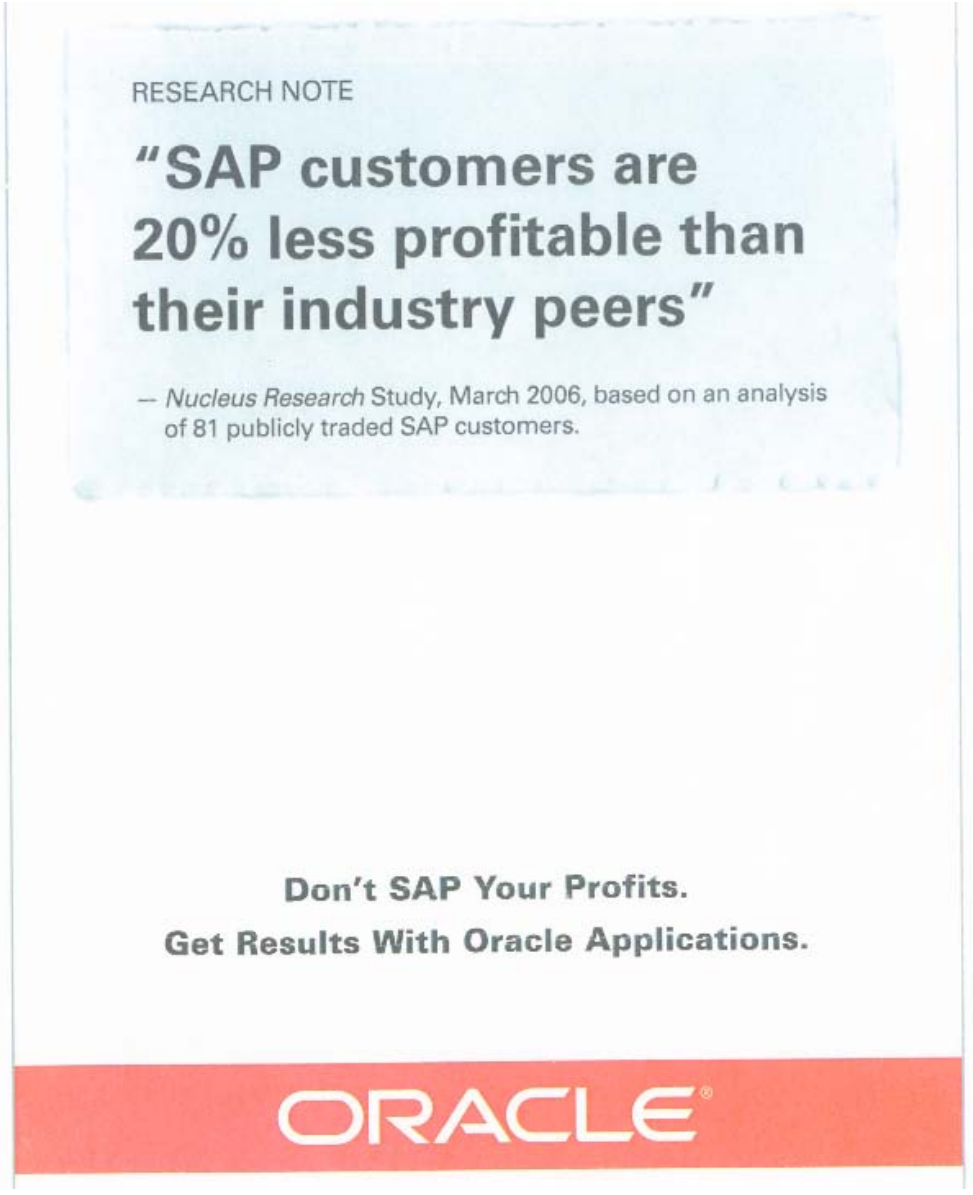
$.82^{10} = 0.137.$

Models, Parameters, Estimates...

In general we talk about unknown quantities using the language of probability... and the following steps:

- Define the random variables of interest
- Define a model (or probability distribution) that describes the behavior of the RV of interest
- Based on the data available, we estimate the parameters defining the model
- We are now ready to describe possible scenarios, generate predictions, make decisions, evaluate risk, etc...

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

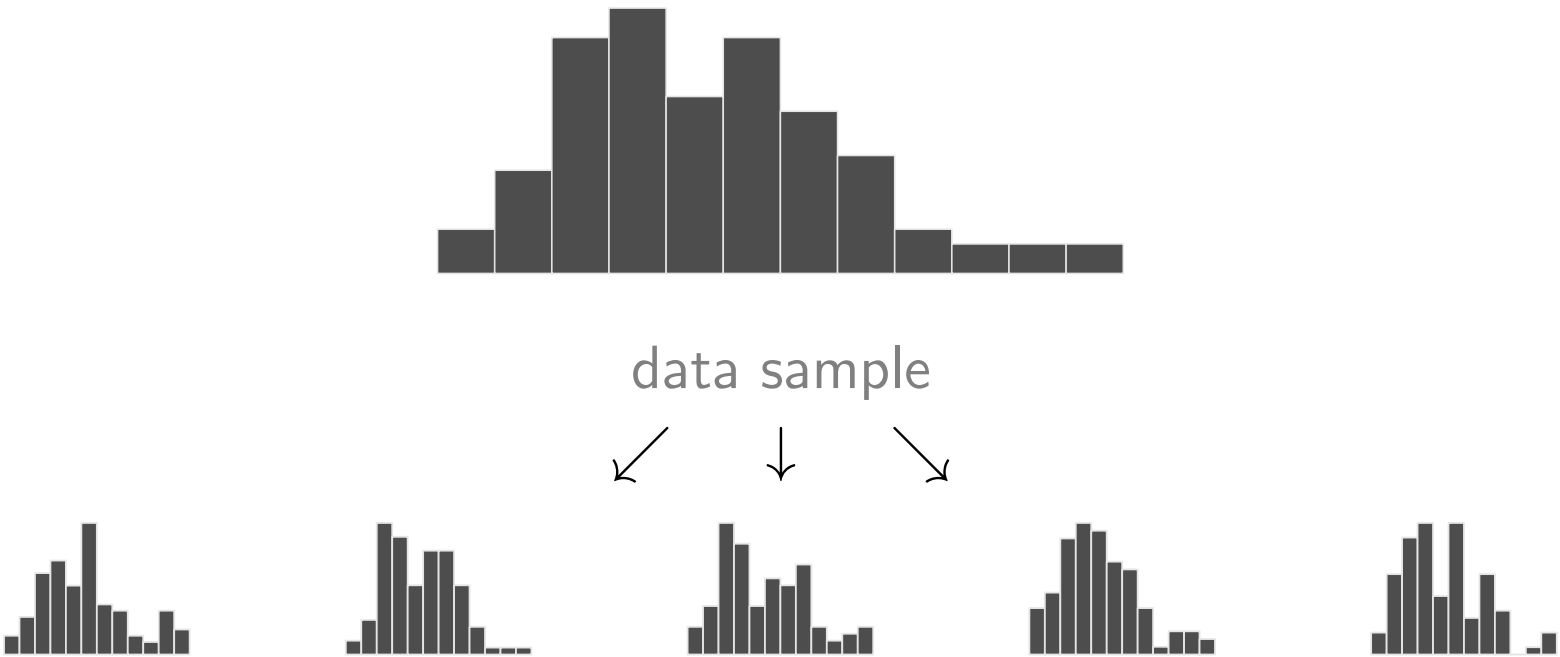
- Well, $\frac{0.12}{0.15} \approx 0.8$! I guess the add is correct, right?
- Not so fast...

Oracle vs. SAP

- Let's assume the sample we have is a good representation of the "population" of firms that use SAP...
- What if we have observed a different sample of size 81?

Oracle vs. SAP

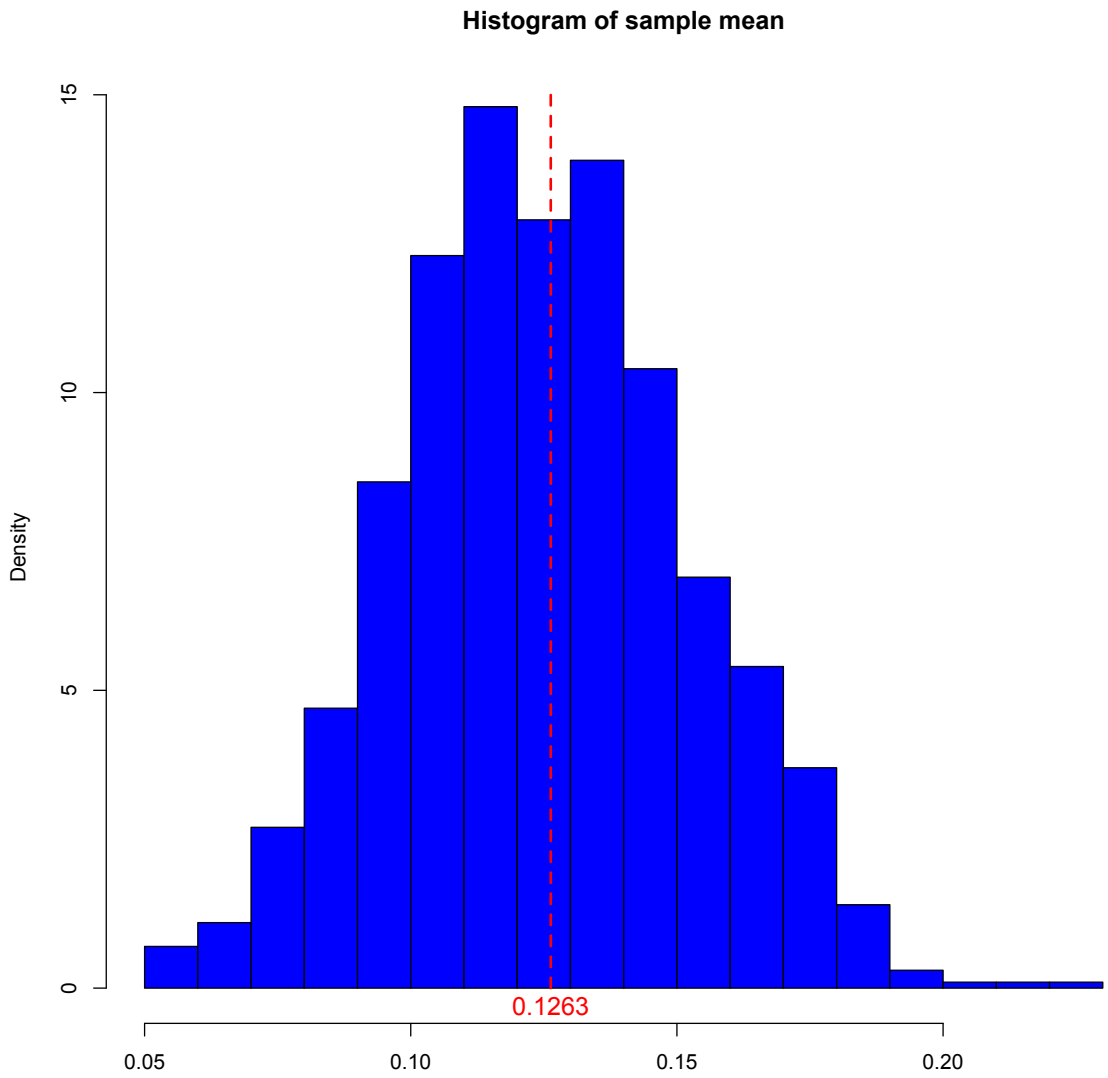
- Selecting a random, with replacement, from the original 81 samples I get a new $\bar{X} = 0.09...$ I do it again, and I get $\bar{X} = 0.155...$ and again $\bar{X} = 0.132...$



Oracle vs. SAP

- After doing this 1000 times... here's the histogram of $\bar{X}$...

  Now, what do you think about the add?



Sampling Distribution of Sample Mean

Consider the mean for an *iid* sample of n observations of a random variable $\{X_1, \ldots, X_n\}$

If X is normal, then

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

.

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

Sampling Distribution of Sample Mean

- The sampling distribution of $\bar{X}$ describes how our estimate would vary over different datasets of the same size n
- It provides us with a vehicle to evaluate the uncertainty associated with our estimate of the mean...
- It turns out that s^2 is a good proxy for σ^2 so that we can approximate the sampling distribution by

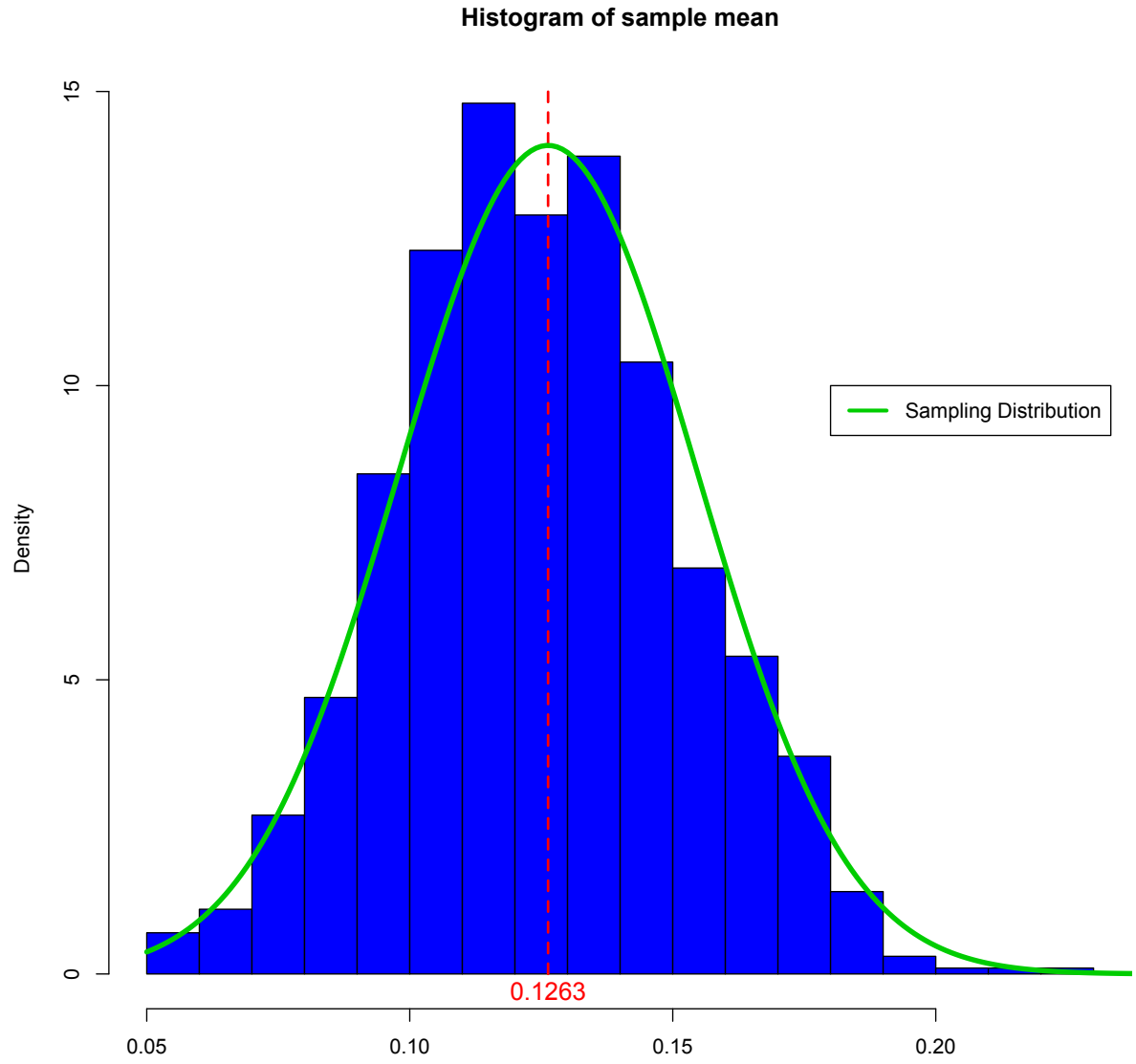
$$\bar{X} \sim N\left(\mu, \frac{s^2}{n}\right)$$

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

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

Back to the Oracle vs. SAP example

Back to our simulation...



Confidence Intervals

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

so...

$$(\bar{X} - \mu) \sim N\left(0, s^2_{\bar{X}}\right)$$

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

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

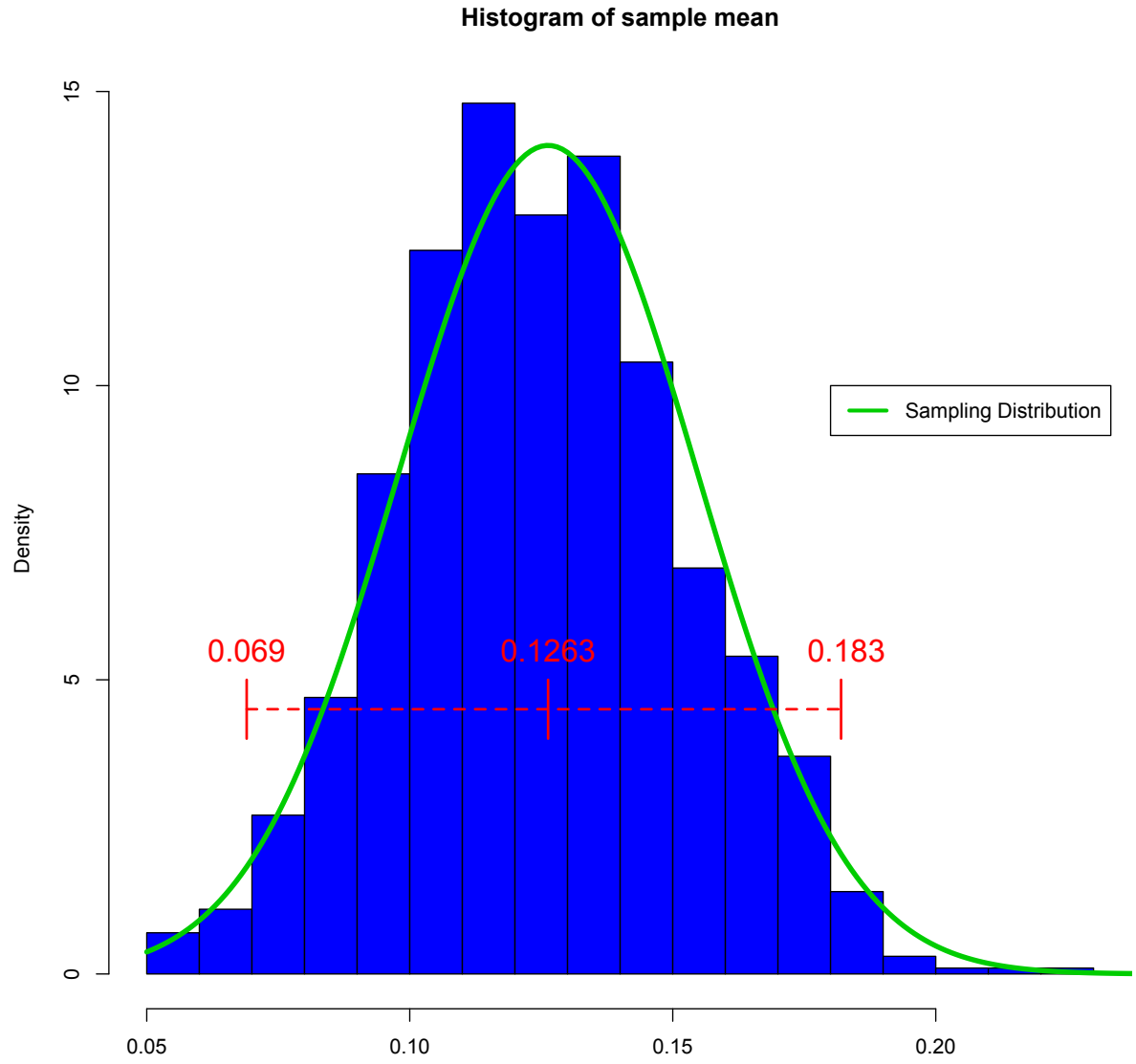
$$= \left[0.1263 - 2 \times \sqrt{\frac{0.065}{81}}; 0.1263 + 2 \times \sqrt{\frac{0.065}{81}}\right]$$

$$= [0.069; 0.183]$$

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

Back to the Oracle vs. SAP example

Back to our simulation...



Let's revisit the US stock market example from before...

Let's run a simulation based on our results...

Estimating Proportions...

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:

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

Estimating Proportions...

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 yes or no with yes $\approx Y_i = 1$ and no $\approx Y_i = 0$, where p is the true population proportion of yes.

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

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

- The same idea applies when comparing means or proportions

The Importance of Considering and Reporting Uncertainty

In 1997 the Red River flooded Grand Forks, ND overtopping its levees with a 54-feet crest. 75% of the homes in the city were damaged or destroyed!

It was predicted that the rain and the spring melt would lead to a 49-feet crest of the river. The levees were 51-feet high.

The Water Services of North Dakota had explicitly avoided communicating the uncertainty in their forecasts as they were afraid the public would loose confidence in their abilities to predict such events.

The Importance of Considering and Reporting Uncertainty

It turns out the prediction interval for the flood was $49ft \pm 9ft$ leading to a 35% probability of the levees being overtopped!!

Should we take the point prediction ($49ft$) or the interval as an input for a decision problem?

In general, the distribution of potential outcomes are very relevant to help us make a decision

The Importance of Considering and Reporting Uncertainty

The answer seems obvious in this example (and it is!)... however, you see these things happening all the time as people tend to underplay uncertainty in many situations!

"Why do people not give intervals? Because they are embarrassed!"

Jan Hatzius, Goldman Sachs economists talking about economic forecasts...

Don't make this mistake! Intervals are your friend and will lead to better decisions!