Section 1.4: Learning from data

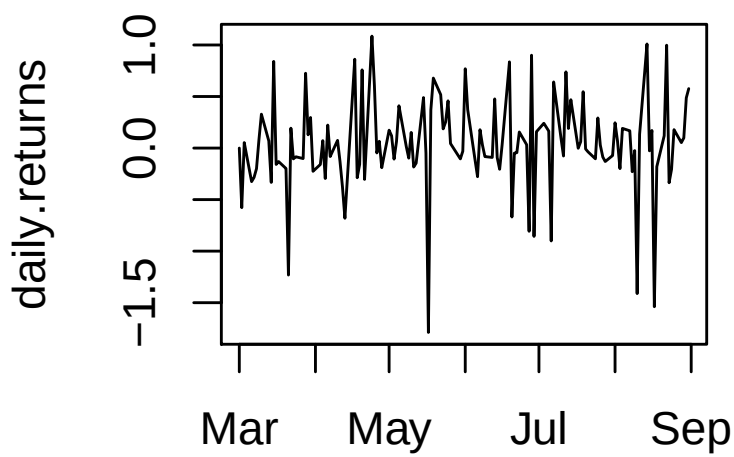
Jared S. Murray
The University of Texas at Austin
McCombs School of Business
Suggested reading: OpenIntro Statistics, Chapter 4.1, 4.2, 4.4, 5.3

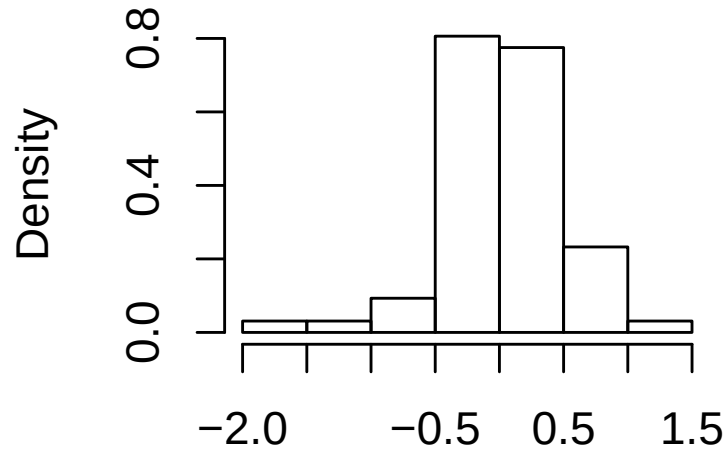
A First Modeling Exercise

- I have US$ 1,000 invested in the SP500. 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 it is that I will lose more than 2% of my money by the end of tomorrow's trading session.
- What should I do?

SP500 - Data

Daily percent returns on the SP500 for Mar 01, 2017 - Sep 01, 2017:





Data are on the website: `https://jaredsmurray.github.io/sta371g_f17/data/sp500_mar-1-17_to_sep-1-17.csv`

- As a first modeling decision, let's call the random variable associated with daily returns on the SP500 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 hand (n=129 observations)...

- Let's assume that each observation in the sample $\{x_1, x_2, x_3, \ldots, x_n\}$ is independent and distributed according to the model above, i.e., $x_i \sim N(\mu, \sigma^2)$
- 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)

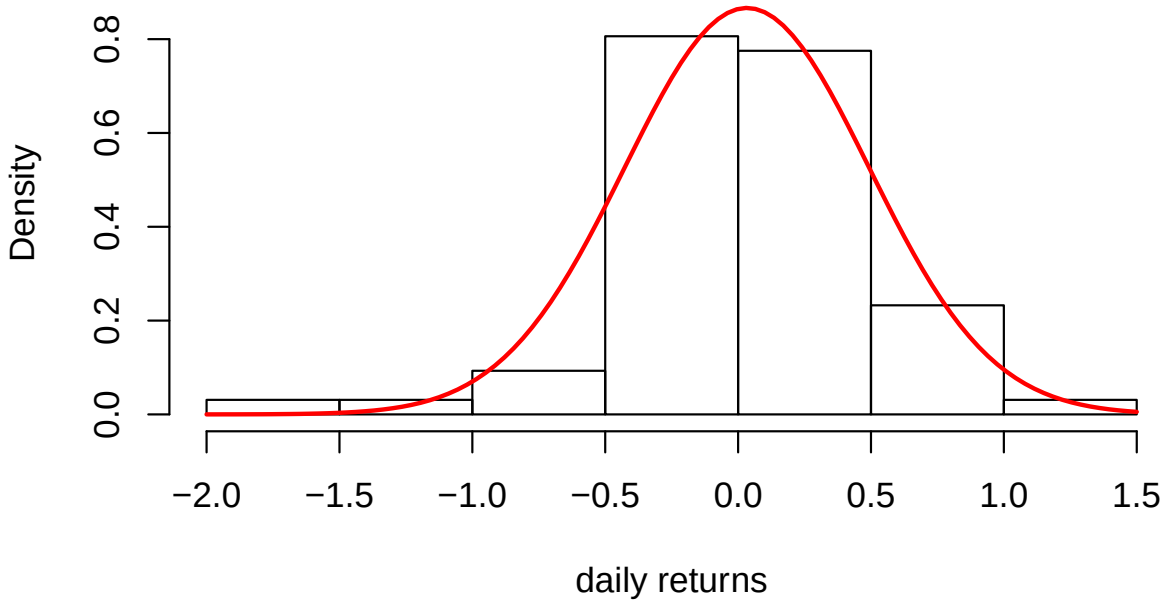
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

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

For the SP500 data in hand, $\bar{X} = 0.03$ and $s^2 = 0.21$

```
xbar = mean(sp500$daily.returns)
s2   = var(sp500$daily.returns)
s    = sd(sp500$daily.returns)
print(xbar)

## [1] 0.03302145

print(s2)

## [1] 0.2119474

print(s)

## [1] 0.4603774
```

For the SP500 data in hand, $\bar{X} = 0.03$ and $s^2 = 0.21$

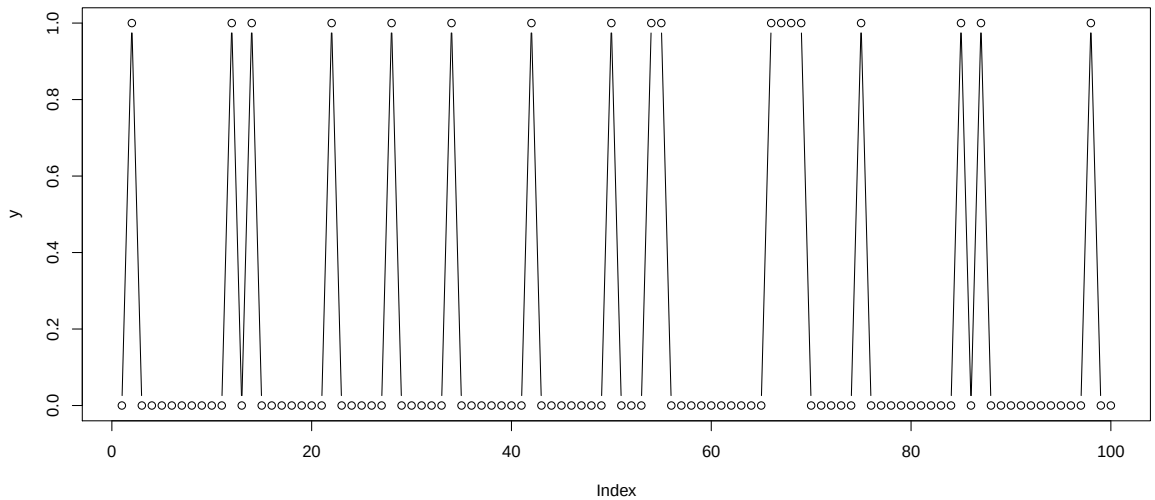


- The red line represents our "model", i.e., the normal distribution with mean and variance given by the estimated quantities $\bar{X}$ and s^2.
- What is $Pr(X < -2)$?

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 independent with the same probability p...

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
$p = \Pr(Y_i = 1) \approx 0.18$.

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

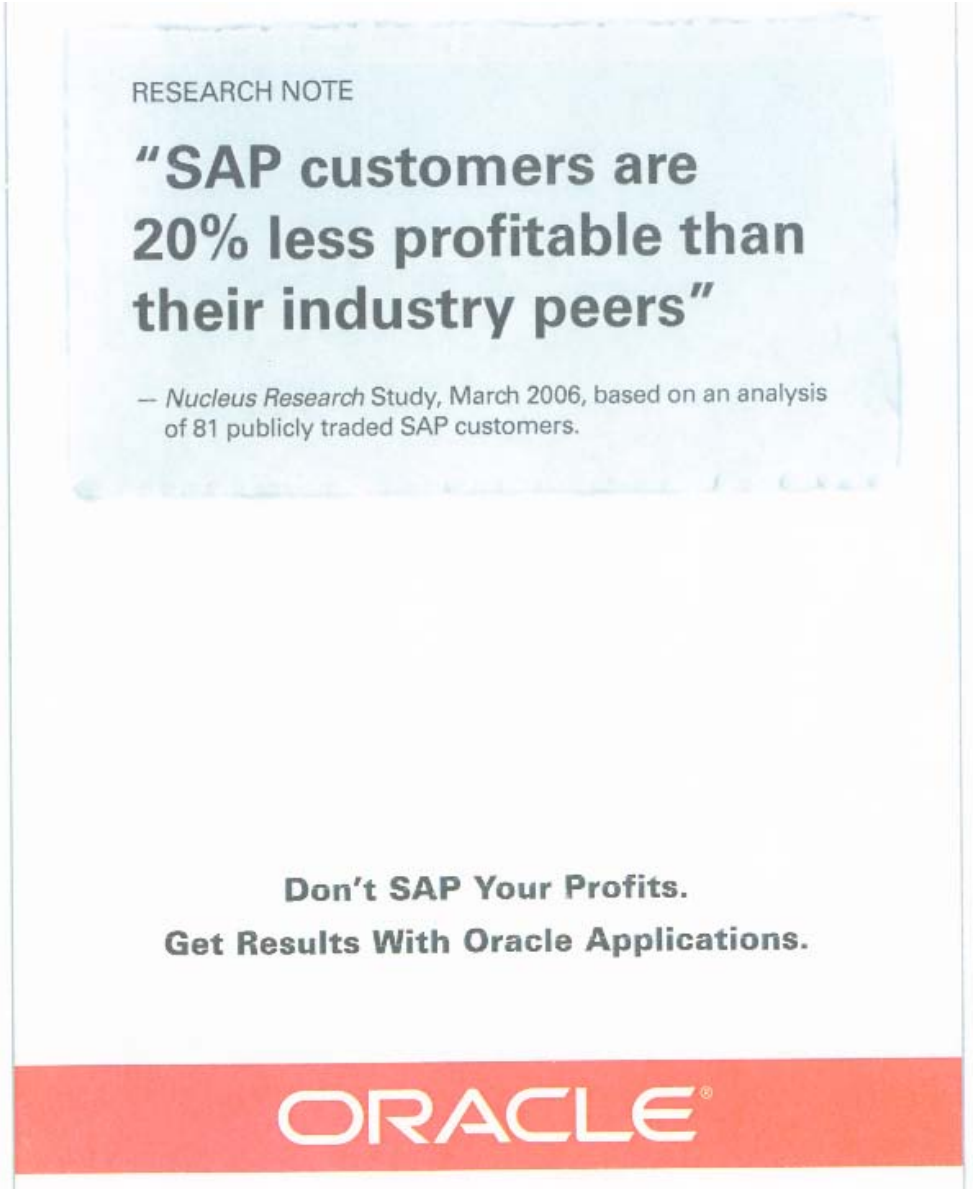
$.82^{10} = 0.137$.

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 (almost) 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 ad?
- We have a dataset of 81 firms that use SAP...
- The industry average 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	0.038

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

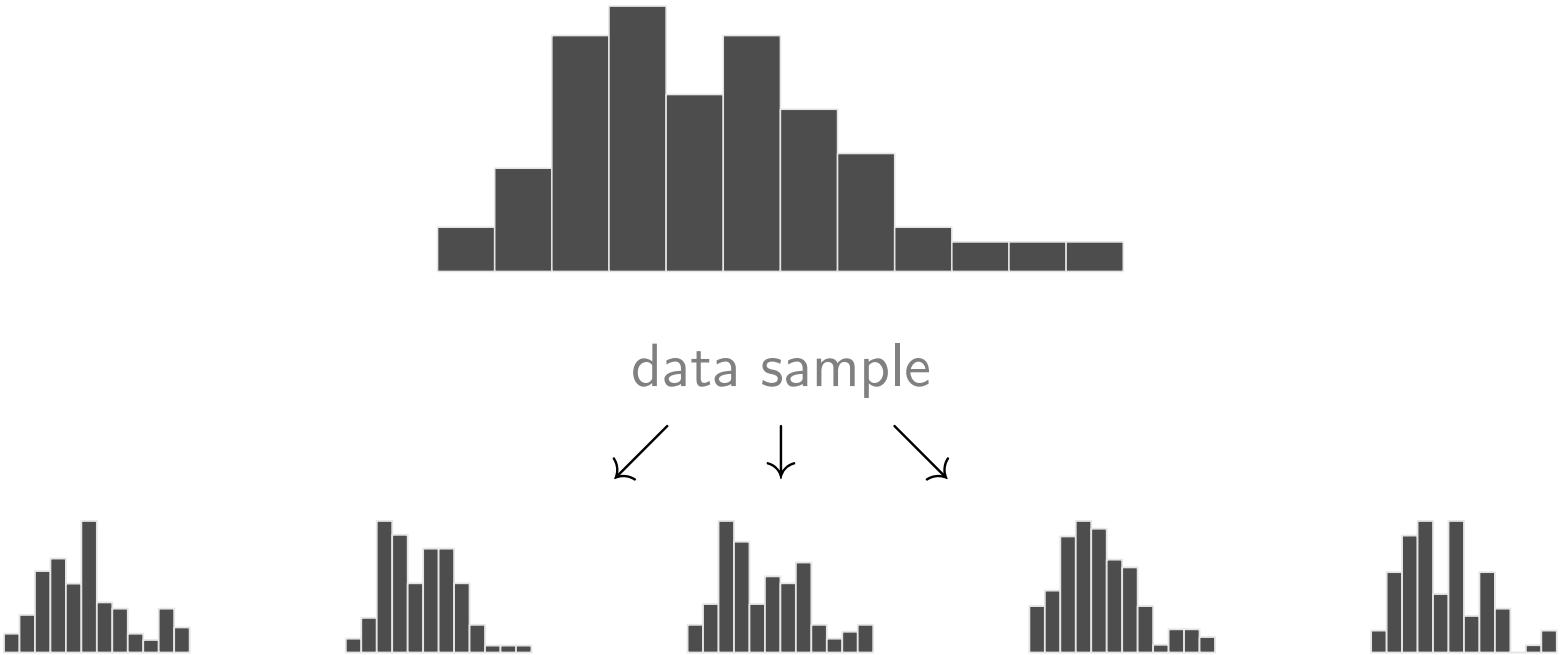
Oracle vs. SAP

- What if we have observed a different sample of size 81? Would the sample mean have been different?

- Let's assume the sample we have is a good representation of the "population" of firms that use SAP...

Oracle vs. SAP

- Sampling 81 observations (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 $\bar{X} = 0.132...$



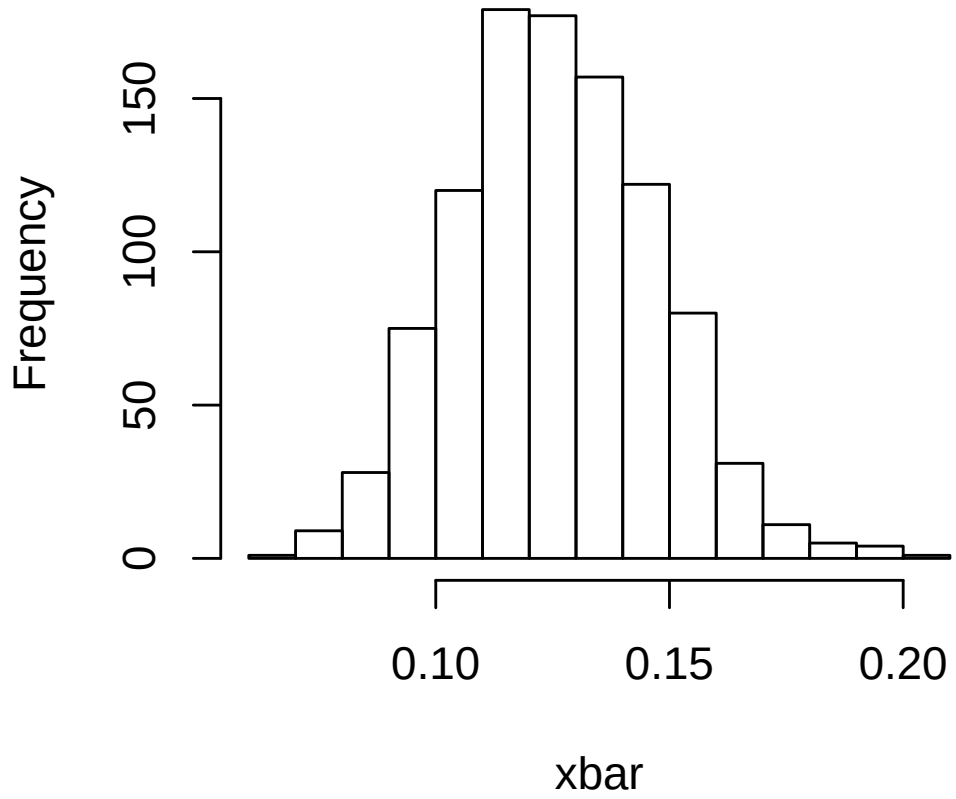
Oracle vs. SAP

This procedure is called "bootstrapping", and it's easy in R: After loading the data into a data frame (or a "tibble") named sap:

```
library(mosaic)
print(sap)

## # A tibble: 81 x 1
##          roe
##        <dbl>
## 1 0.25922835
## 2 0.03313982
## 3 0.01278884
## # ... with 78 more rows

sap.boot = do(1000) * {
    mean(~roe, data = resample(sap))
}
```

Oracle vs. SAP

- After 1,000 samples here's the histogram of $\bar{X}$... Now, what do you think about the ad?



Note: $\Pr(\bar{X} > 0.15) \approx 0.13$

Sampling Distribution of the Sample Mean

What's going on here?

- We're simulating (approximately) the *sampling distribution* of the sample mean, i.e., the probability distribution of $\bar{X}$ - how does $\bar{X}$ vary over datasets of size n?
- The sampling distribution quanitifies uncertainty in our estimates - $\bar{X} \neq \mu$, but how wrong might we be?
- We have one more important tool for estimating sampling distributions

Central Limit Theorem

Consider the mean for a sample of n independent observations of a random variable: $\{X_1, X_2, \ldots, X_n\}$

Suppose that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

- $E(\bar{X}) = \frac{1}{n} \sum E(X_i) = \mu$
- $Var(\bar{X}) = Var\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \sum Var\left(X_i\right) = \frac{\sigma^2}{n}$

For large n,

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

(approximately)

Sampling Distribution of Sample Mean

- It turns out that s^2 is a good proxy for σ^2, so 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

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

Sampling Distribution of Sample Mean

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

- $\bar{X}$ 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!

Keep track of your s's: s^2 and $s^2_{\bar{X}}$

In our $N(\mu, \sigma^2)$ model...

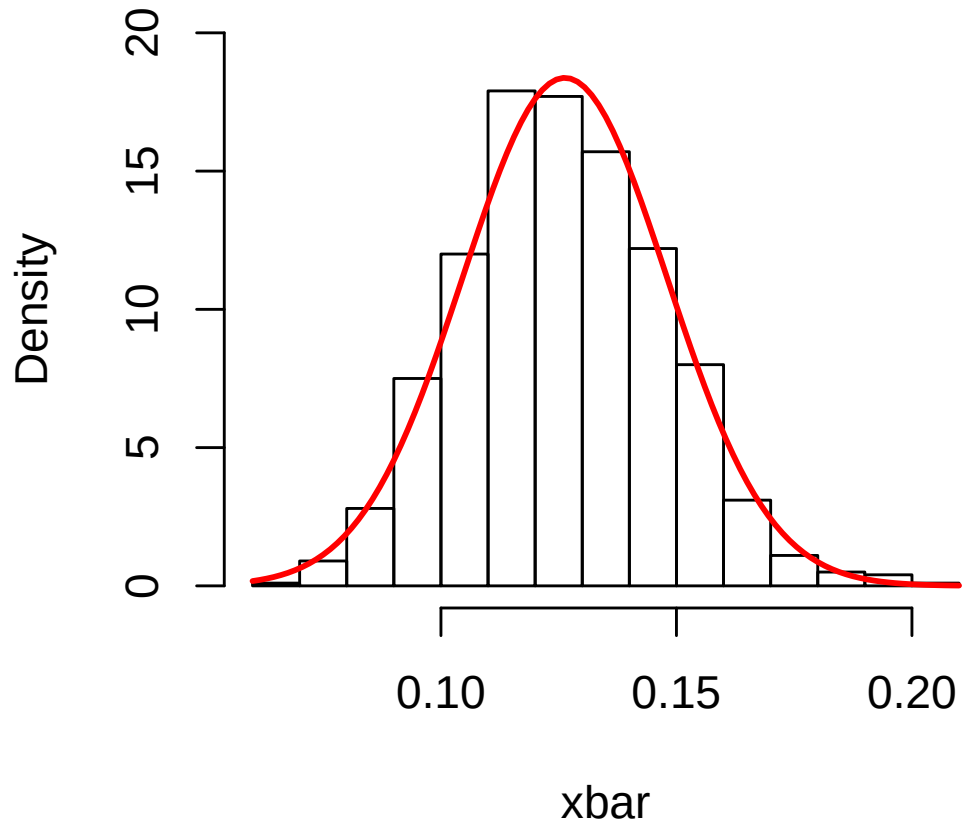
- s^2 is an estimate of the observation variance σ^2, or **how close a single observation tends to be to its expected value** μ.
- $s^2_{\bar{X}}$ is an estimate of the **sample mean ($\bar{X}$)'s variance σ^2/n, or how close the sample mean of n observations tends to be to its expected value** μ

Roughly:

- s^2 estimates uncertainty in future observations if we knew μ
- $s^2_{\bar{X}}$ estimates uncertainty about μ

Back to the Oracle vs. SAP example

Back to our simulation...



The two approximations to the sampling distribution are very close.

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% confidence interval for** μ**.** In general, you can think of this as a set of **plausible values** for μ

Oracle vs. SAP example... one more time

In this example, $\bar{X} = 0.1263$, $s^2 = 0.038$ and $n = 81$... therefore, $s^2_{\bar{X}} = \frac{0.038}{81}$ so, the 95% confidence interval for the ROE of SAP firms is

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

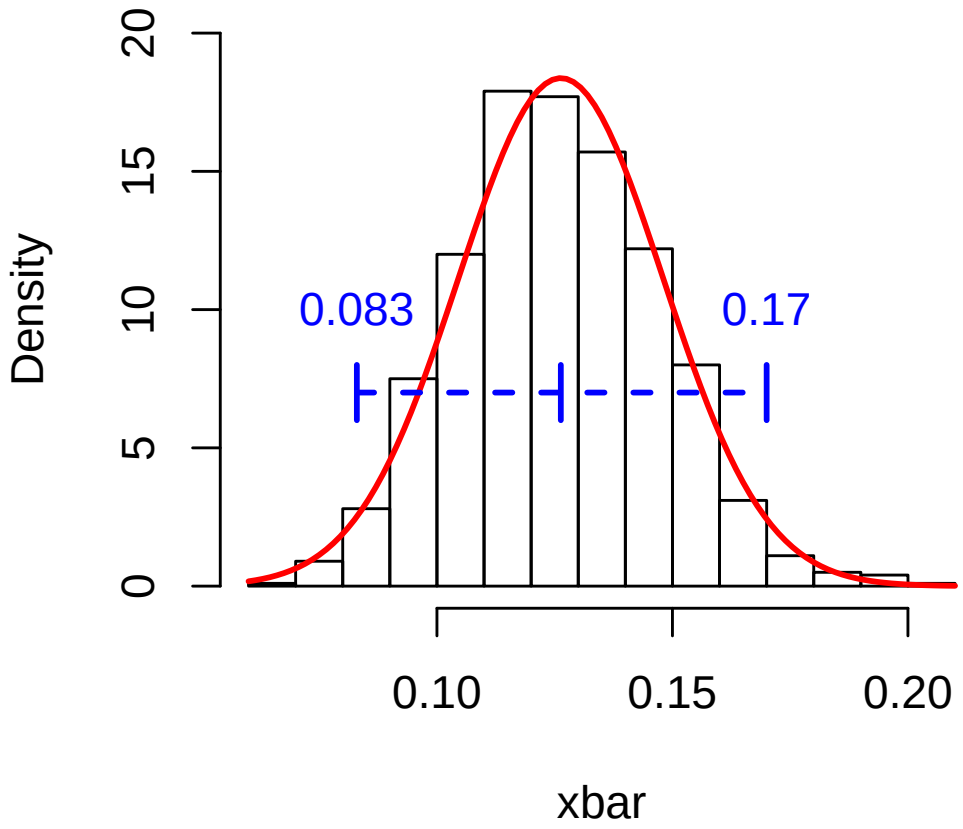
$$= \left[0.1263 - 2 \times \sqrt{\frac{0.038}{81}}; 0.1263 + 2 \times \sqrt{\frac{0.038}{81}}\right]$$

$$= [0.083; 0.170]$$

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

Back to the Oracle vs. SAP example

Back to our simulation...



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. (note: the sample proportion is just the sample mean of a binary r.v.)

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...

Suppose we take a relatively small random sample from a large population and ask each respondent a question, with yes corresponding to $Y_i = 1$ and no to $Y_i = 0$. Let p be the true population proportion of yes's.

Suppose $n = 1000$, and $p = .5$ so $\hat{p} \approx 0.5$ (Remember that $Var(Y_i) = p(1-p)$ is largest when $p = 0.5$)

Then,

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

The standard error is .0158 so that half of a 95% CI (the "$\pm$") is .0316, or about $\pm$ 3%.

The Bottom Line...

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

- The "standard error" measures the uncertainty of an estimate
- For most parameters a good "95% Confidence Interval" is

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

- 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 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 topped!!

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

In general, the distribution of possible outcomes (not a single prediction/estimate) is relevant for decisionmaking.

The Importance of Considering and Reporting Uncertainty

The answer seems obvious in this example (and it is!)... however, people tend to underplay uncertainty in many situations.

"Why do people not give intervals? Because they are embarrassed!"
Jan Hatzius, Goldman Sachs chief economist, talking about economic forecasts...

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