TEXAS McCombs

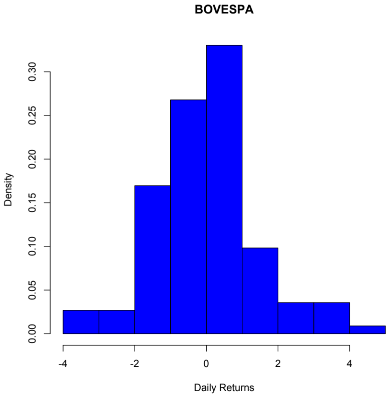
Modeling, estimation, and testing

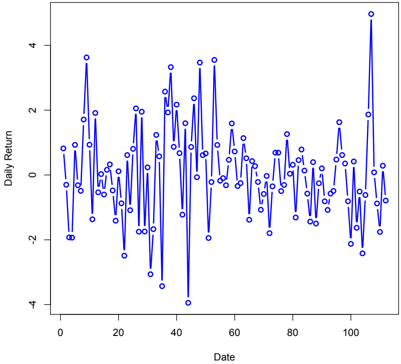
David Puelz

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

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

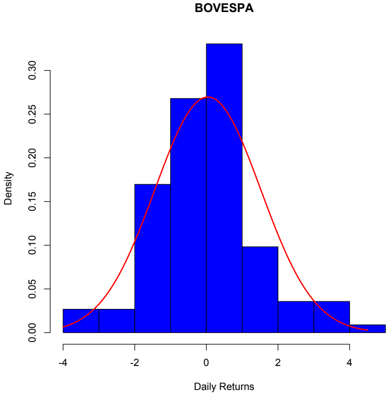
- Question: What are the values of μ and σ^2 ?
- We need to estimate these values from our sample (n=113 observations)...

- 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., $x_i \sim N(\mu, \sigma^2)$
- The typical 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,



$\bar{X} = 0.04$ 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 $Pr(X < -3)$?

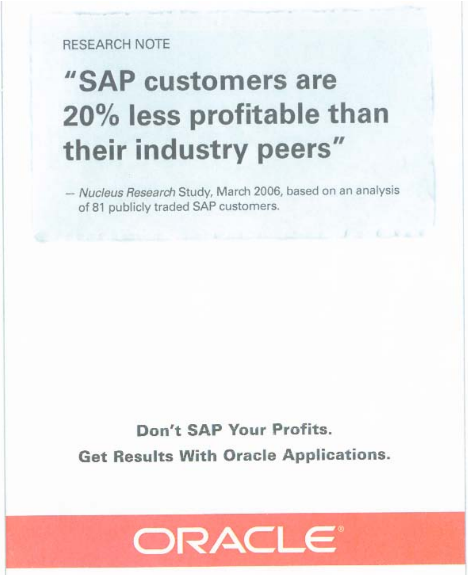
Models, parameters, estimates...

In general we talk about unknown quantities using the language of probability, i.e., probability models. The steps to follow are:

- Define the random variable(s) of interest
- Specify a model (aka probability distribution) that describes the behavior of the RV of interest
- Based on the data available, estimate the parameters defining the model

Now, we are now ready to describe possible scenarios, generate predictions, make decisions, evaluate risk, etc...

Oracle vs. SAP example (understanding variation)



RESEARCH NOTE

"SAP customers are 20% less profitable than their industry peers"

– *Nucleus Research* Study, March 2006, based on an analysis of 81 publicly traded SAP customers.

Don't SAP Your Profits.

Get Results With Oracle Applications.

ORACLE®

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)
- from the data:

	$\bar{X}$	s
SAP firms	0.1263	0.25

- 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?
- Would the average be the same?
- How different?? Enter the LeBron James of statistics...

Sampling distribution of sample mean

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

It turns out

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

.

This is called the sampling distribution of the mean... the fact the average follows a normal distribution is what the book calls "the central limit theorem" (a.k.a. Lebron James)

Standard error

- 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

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

Standard error

- The standard error measures the dispersion on the sample average ($\bar{X}$) around the underlying population mean (μ)

- Looking at the formula:

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

  we see that

  - the sample average will cluster "tightly" around the population mean as the size of the sample gets larger (as n grows)

  - the sample average will cluster "less tightly" around the population mean when the underlying population in more spread out (when s is bigger).

Standard deviation (s) vs. standard error ($s_{\bar{X}}$)

- The standard deviation measures the dispersion within the sample
- The standard error measures the dispersion of the average across samples of the population!
- Put simply, the standard error is the standard deviation of the sampling distribution!

Confidence intervals

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

so...

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

- 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, $\bar{X} = 0.1263$, $s = 0.25$ and $n = 81$... therefore, $s_{\bar{X}} = \sqrt{\frac{0.25^2}{81}} = 0.028$ so, the 95% confidence interval for the ROE of SAP firms is

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

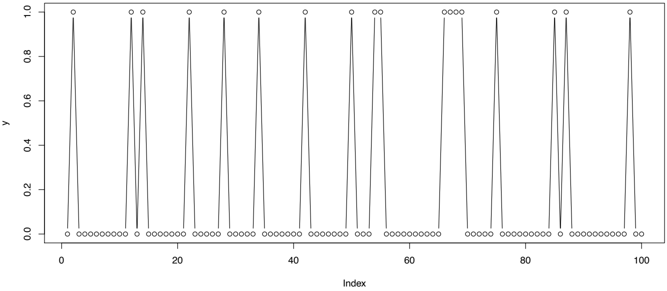
$$= [0.1263 - 2 \times 0.028; 0.1263 + 2 \times 0.028]$$

$$= [0.069; 0.183]$$

– Is 0.15 a plausible value? Yes!

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 means 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 probability!

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

$.82^{10} = 0.137$.

Standard error of 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:

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

Standard error of 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: Another example...

(Read chapter 10 of "Naked Statistics" if you have a chance)

If we take a relatively small random sample from a large population and ask each respondent for an answer 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%.

(Sound familiar?)

Example: Salary discrimination

Say we are concerned with potential salary discrimination between males and females in the banking industry... To study this issue, we get a sample of salaries for both 100 males and 150 females from multiple banks in Chicago. Here is a summary of the data:

	average	std. deviation
males	150k	30k
females	143k	15k

What do we conclude? Is there a difference for sure?

Example: Salary discrimination

Let's compute the confidence intervals:

males:

$$(150 - 2 \times \sqrt{\frac{30^2}{100}}; 150 + 2 \times \sqrt{\frac{30^2}{100}}) = (144; 156)$$

females:

$$(143 - 2 \times \sqrt{\frac{15^2}{150}}; 143 + 2 \times \sqrt{\frac{15^2}{150}}) = (140.55; 145.45)$$

How about now, what do we conclude?

Example: Google search algorithm

Google is testing a couple of modifications in its search algorithms... they experiment with 2,500 searches and check how often the result was defined as a "success." Here's the data from this experiment:

Algorithm	current	mod 1	mod 2
success	1755	1850	1760
failure	745	650	740

The probability of success is estimated to be $\hat{p} = 0.702$ for the current algorithm, $\hat{p}_A = 0.74$ for modification (A) and $\hat{p}_B = 0.704$ for modification (B) .

Are the modifications better for sure?

Example: Google search algorithm

Let's compute the confidence intervals and check if these modifications are REALLY working...

current:

$$\left(.702 - 2 \times \sqrt{\frac{.702 * (1 - .702)}{2500}}; .702 + 2 \times \sqrt{\frac{.702 * (1 - .702)}{2500}}\right) = (0.683; 0.720)$$

mod (A):

$$\left(.740 - 2 \times \sqrt{\frac{.740 * (1 - .740)}{2500}}; .740 + 2 \times \sqrt{\frac{.740 * (1 - .740)}{2500}}\right) = (0.723; 0.758)$$

mod (B):

$$\left(.704 - 2 \times \sqrt{\frac{.704 * (1 - .704)}{2500}}; .704 + 2 \times \sqrt{\frac{.704 * (1 - .704)}{2500}}\right) = (0.686; 0.722)$$

What do we conclude?

Standard error for the difference in means

It turns out there is a more precise way to address these comparisons problems (for two groups)...

We can compute the *standard error for the difference in means:*

$$s_{(\bar{X}_a-\bar{X}_b)} = \sqrt{\frac{s^2_{X_a}}{n_a} + \frac{s^2_{X_b}}{n_b}}$$

or, for the *difference in proportions*

$$s_{(\hat{p}_a-\hat{p}_b)} = \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}$$

Confidence interval for the difference in means

We can then compute the
confidence interval for the difference in means:

$$(\bar{X}_a - \bar{X}_b) \pm 2 \times s_{(\bar{X}_a - \bar{X}_b)}$$

or, the
confidence interval for the difference in proportions

$$(\hat{p}_a - \hat{p}_b) \pm 2 \times s_{(\hat{p}_a - \hat{p}_b)}$$

Let's revisit the examples... salary discrimination

$$s_{(\bar{X}_{males} - \bar{X}_{females})} = \sqrt{\frac{30^2}{100} + \frac{15^2}{150}} = 3.24$$

so that the confidence interval for the difference in means is:

$$(150 - 143) \pm 2 \times 3.24 = (0.519; 13.48)$$

What is the conclusion now?

Let's revisit the example... Google search

Let's look at the difference between the current algorithm and modification B...

$$s_{(\hat{p}_{current} - \hat{p}_{new})} = \sqrt{\frac{0.702 * 0.298}{2500} + \frac{0.704 * 0.296}{2500}} = 0.0129$$

so that the confidence interval for the difference in means is:

$$(0.702 - 0.704) \pm 2 \times 0.0129 = (-0.0278; 0.0238)$$

What is the conclusion now?

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

Testing

Suppose we want to assess whether or not μ equals a proposed value μ^0. This is called hypothesis testing.

Formally we test the null hypothesis:

$H_0: \ \mu = \mu^0$

vs. the alternative

$H_1: \ \mu \neq \mu^0$

Testing

That are 2 ways we can think about testing:

1. Building a test statistic... the t-stat,

$$t = \frac{\bar{X} - \mu^0}{s_{\bar{X}}}$$

   This quantity measures how many standard deviations the estimate ($\bar{X}$) from the proposed value (μ^0).

   If the absolute value of t is greater than 2, we need to worry (why?)... we reject the hypothesis.

Testing

2. Looking at the confidence interval. If the proposed value is outside the confidence interval you reject the hypothesis.

   Notice that this is equivalent to the t-stat. An absolute value for t greater than 2 implies that the proposed value is outside the confidence interval... therefore reject.

   This is my preferred approach for the testing problem. You can't go wrong by using the confidence interval!

Testing (proportions)

- The same idea applies to proportions... we can compute the t-stat testing the hypothesis that the true proportion equals p^0

$$t = \frac{\hat{p} - p^0}{s_{\hat{p}}}$$

  Again, if the absolute value of t is greater than 2,
  we reject the hypothesis.

- As always, the confidence interval provides you with the same (and more!) information.

(Note: In the proportion case, this test is sometimes called a z-test)

Testing (differences)

- For testing the difference in means:

$$t = \frac{(\bar{X}_a - \bar{X}_b) - d^0}{s_{(\bar{X}_a - \bar{X}_b)}}$$

- For testing a difference in proportions:

$$t = \frac{(\hat{p}_a - \hat{p}_b) - d^0}{s_{(\hat{p}_a - \hat{p}_b)}}$$

In both cases d^0 is the proposed value for the difference (we often think of zero here... why?)

Again, if the absolute value of t is greater than 2,
we reject the hypothesis.

(Note: In the proportion case, this test is sometimes called a z-test)

Testing (examples)

Let's recap by revisiting some examples:

- What hypothesis were we interested in the Oracle vs. SAP example? Use a t-stat to test it...

- Use the t-stat to determine whether or not males are paid more than females in the Chicago banking industry

- What does the t-stat tells you about Google's new search algorithm?