Section 2: Estimation, Confidence Intervals and Testing Hypothesis

Tengyuan Liang, Chicago Booth

https://tyliang.github.io/BUS41000/

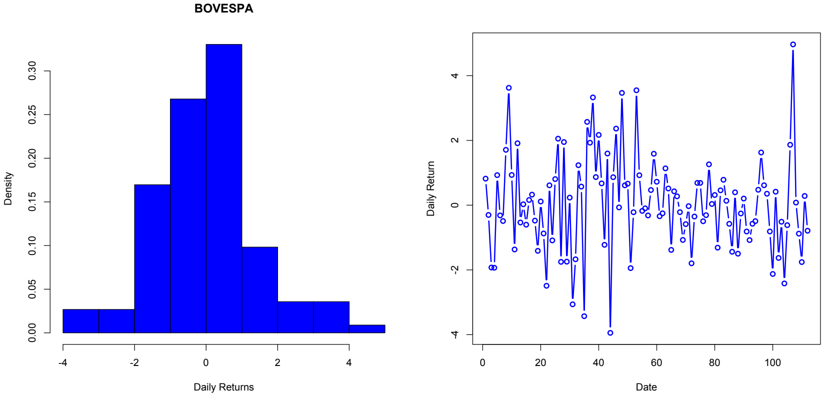
Suggested Reading:
Naked Statistics, Chapters 7, 8, 9 and 10
OpenIntro Statistics, Chapters 4, 5 and 6

Statistical Modeling: Parameters and Estimates

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

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

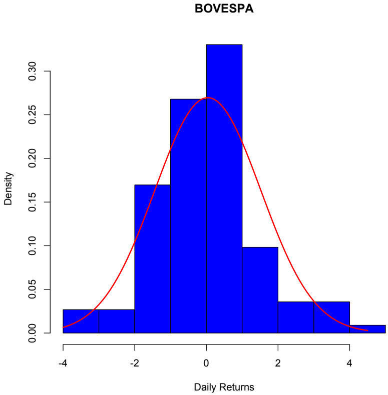
- Question: What are the values of μ and σ^2 ?
- We need to estimate these values from the sample in hands (n=113 observations)...

- 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 $s^2 = 2.19$

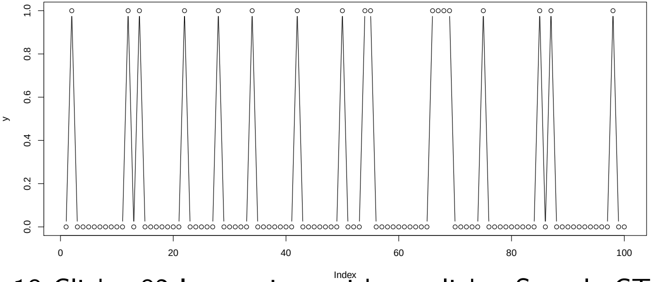
- The red line represents our "model," i.e., the normal distribution with mean and variance given by the estimated quantities $\bar{X}$ and s^2.
- What is $Pr(X < -3)$?

Estimating Proportions: A/B Testing & CTR

Scenario: You are a Product Manager at a tech startup launching a new **Instagram Ad**. You pilot the ad to 100 users.

- $Y_i = 1$: User clicked (Conversion)
- $Y_i = 0$: User scrolled past (No Interest)

Objective: Estimate the true **Click-Through Rate (CTR)** to decide if we should scale the campaign budget.



Data: 18 Clicks, 82 Impressions without clicks. Sample CTR = 18%.

Modeling User Behavior

Modeling Assumption: We model user clicks as **Independent and Identically Distributed (i.i.d.)** Bernoulli trials.

- **Independence:** One user's click doesn't affect another's. (Plausible?)
- **Identical:** Every user has the same probability p of clicking. (Target audience homogeneity)

Prediction: If the true $p \approx 0.18$ (based on our sample), what is the probability that the next 10 users **all ignore** the ad?

$$P(10 \text{ failures}) = (1 - 0.18)^{10} \approx 0.82^{10} = 0.137$$

There is a $\approx 13.7\%$ chance of a "dry spell" of 10 users, even with a healthy 18% CTR.

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. . .

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?

Building Portfolios

Building Portfolios

- Let's assume we are considering 3 investment opportunities
  1. IBM stocks
  2. ALCOA stocks
  3. Treasury Bonds (T-bill)
- How should we start thinking about this problem?

Building Portfolios

Let's first learn about the characteristics of each option by assuming the following models:

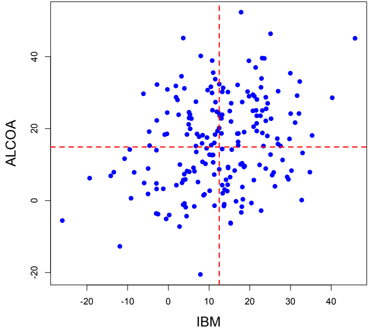
- $IBM \sim N(\mu_I, \sigma_I^2)$
- $ALCOA \sim N(\mu_A, \sigma_A^2)$

and

- The return on the T-bill is 3%

After observing some return data we can came up with estimates for the means and variances describing the behavior of these stocks

Building Portfolios



IBM	ALCOA	T-bill
$\hat{\mu}_I = 12.5$ $\hat{\sigma}_I = 10.5$	$\hat{\mu}_A = 14.9$ $\hat{\sigma}_A = 14.0$	$\mu_{Tbill} = 3$ $\sigma_{Tbill} = 0$

$$corr(IBM, ALCOA) = 0.33$$

Building Portfolios

- How about combining these options? Is that a good idea? Is it good to have all your eggs in the same basket? Why?
- What if I place half of my money in ALCOA and the other half on T-bills...

- Remember that:

$$\begin{aligned} E(aX + bY) &= aE(X) + bE(Y) \\ Var(aX + bY) &= a^2 Var(X) + b^2 Var(Y) + 2ab * Cov(X, Y) \end{aligned}$$

Building Portfolios

- So, by using what we know about the means and variances we get to:

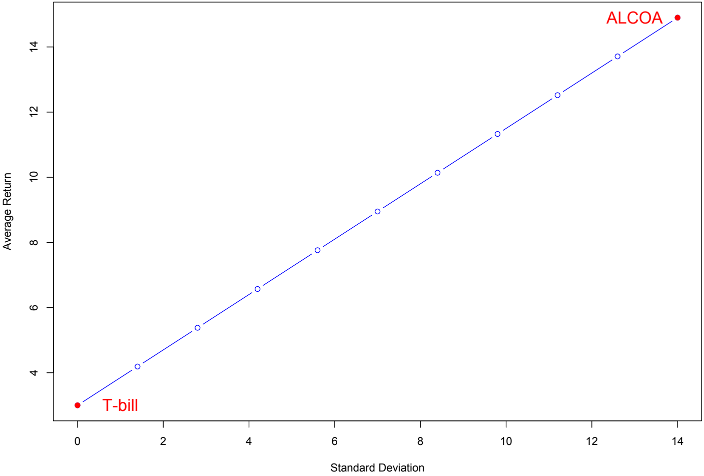
$$\begin{aligned}\hat{\mu}_P &= 0.5\hat{\mu}_A + 0.5\mu_{Tbill} \\ \hat{\sigma}_P^2 &= 0.5^2\hat{\sigma}_A^2 + 0.5^2 * 0 + 2 * 0.5 * 0.5 * 0\end{aligned}$$

- $\hat{\mu}_P$ and $\hat{\sigma}_P^2$ refer to the estimated mean and variance of our portfolio

- What are we assuming here?

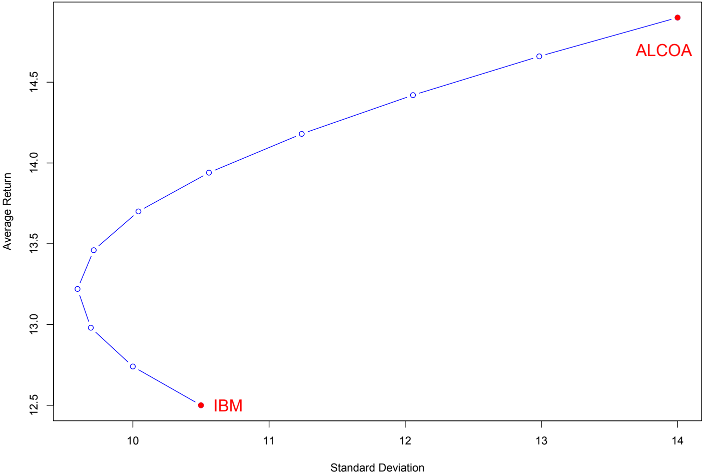
Building Portfolios

- What happens if we change the proportions...



Building Portfolios

- What about investing in IBM and ALCOA?



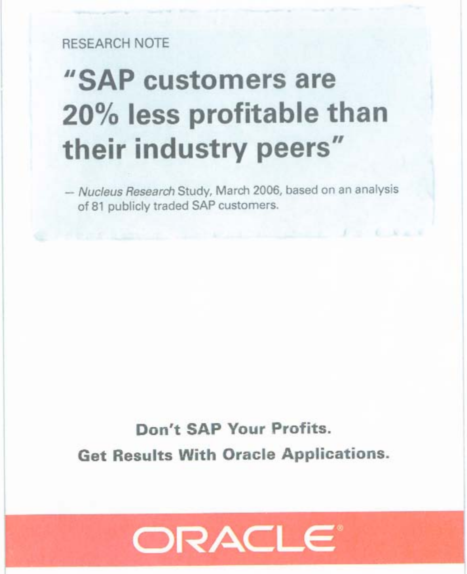
How much more complicated this gets if I am choosing between 100 stocks?

Sampling distribution of an individual

vs.

Sampling distribution of the sample mean

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)
- 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

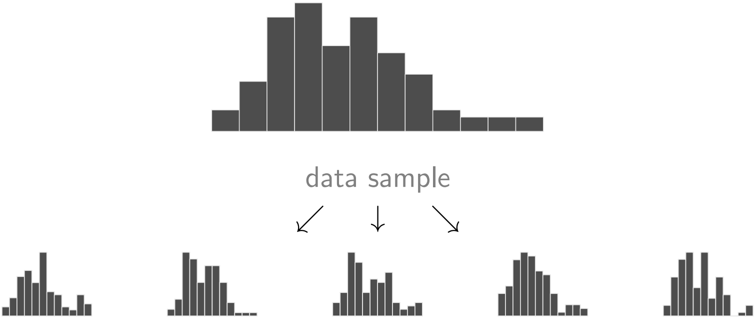
- Well, $\frac{0.12}{0.15} \approx 0.8$! I guess the ad is correct, right?
- Not so fast...

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?

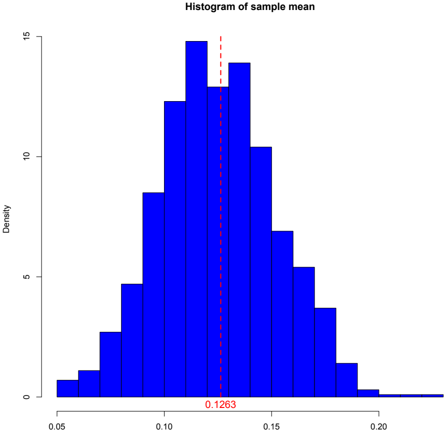
Oracle vs. SAP

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



Oracle vs. SAP

- After doing this 1000 times... here's the histogram of $\bar{X}$... Now, what do you think about the ad?



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

Sampling Distribution of Sample Mean

The approach we numerically obtain the sampling distribution (of the sample mean) is called the Bootstrap.

Bootstrap is a very useful technique in statistics that allow us to obtain the sampling distribution of almost any statistic using random sampling methods.

Sampling Distribution of Sample Mean

Yet another approach: normal approximation, or Central Limit Theorem.

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

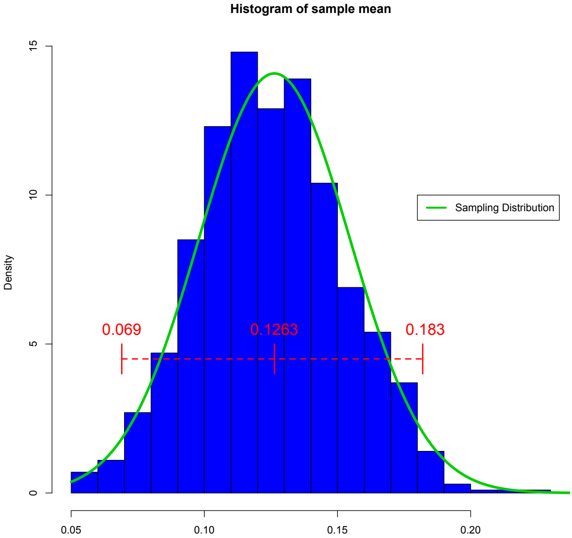
If X is normal, then

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

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

Back to the Oracle vs. SAP example

Back to our simulation. . .



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

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

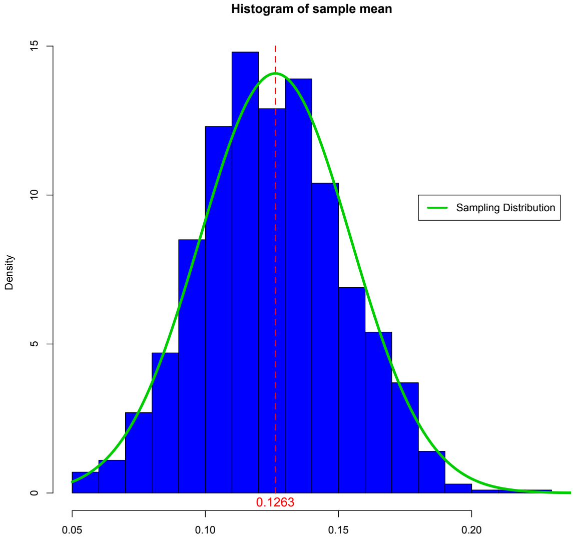
Sampling Distribution of Sample Mean

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

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

Back to the Oracle vs. SAP example

Back to our simulation. . .



Confidence Intervals

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

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

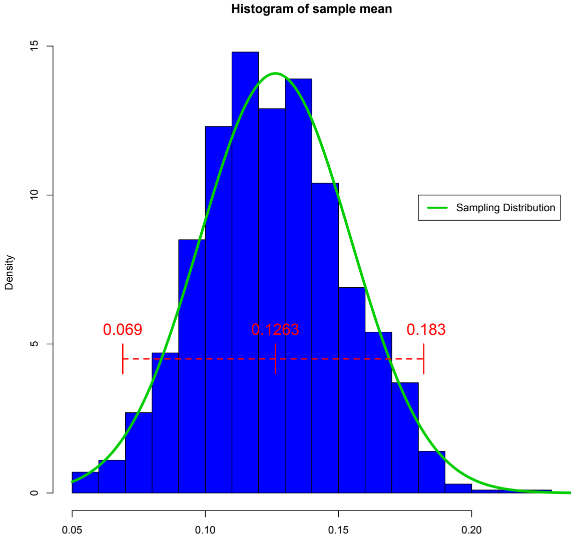
$$= \left[0.1263 - 2 \times \sqrt{\frac{0.065}{81}}; 0.1263 + 2 \times \sqrt{\frac{0.065}{81}}\right]$$

$$= [0.069; 0.183]$$

- 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...

```r
# Generate 1000 parallel worlds, each 90 years of SP500 re
returns = matrix(rnorm(1000*90, mean=11.5, sd=19.5),
                 nrow = 1000, ncol = 90)
x_bar = apply(returns, 1, mean)
se_x = apply(returns, 1, sd)/sqrt(90)

# Volatility of X_bar
sd(x_bar)
```

```
## [1] 2.09836
```

```r
# Our mathmatical formula for s_{X_bar}
19.5/sqrt(90)
```

```
## [1] 2.05548
```

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

```
# coverage of CI
CI = data.frame(CI_lower = x_bar-1.96*se_x,
                CI_upper = x_bar+1.96*se_x,
                Covers_mu = as.logical((x_bar-1.96*se_x<11.
head(CI)
```

```
##    CI_lower CI_upper Covers_mu
## 1  6.289712 15.45795      TRUE
## 2  7.428931 14.17286      TRUE
## 3  7.982620 17.47475      TRUE
## 4  8.564611 16.35973      TRUE
## 5  7.253590 15.05963      TRUE
## 6 11.525648 19.67608     FALSE
```

```
mean(CI$Covers_mu)
```

```
## [1] 0.946
```

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... (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 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?!)

Example: Salary Difference

Say we are concerned with potential salary difference 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 Difference, Naive Approach

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: LLM Human Evaluation (Gemini vs. GPT)

You are evaluating two LLMs for a Customer Support Chatbot. You send 2,500 queries to each and a human panel rates if the answer was “Helpful” (Success).

Model	Baseline (GPT-3.5)	Gemini Pro	GPT-4
Helpful	1755	1850	1760
Not Helpful	745	650	740

The success rate is $\hat{p} = 0.702$ for Baseline, $\hat{p}_{Gemini} = 0.74$ and $\hat{p}_{GPT4} = 0.704$.

Is Gemini TRULY better? Is GPT-4 better than Baseline?

Example: LLM Evaluation, Naive Approach

Let's compute the confidence intervals to see if the "lift" is real...

Baseline:

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

Gemini Pro:

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

GPT-4:

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

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 Difference

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

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 examples... LLM Eval

Let's look at the difference between the Baseline and GPT-4 (small lift)...

$$s_{(\hat{p}_{Base} - \hat{p}_{GPT4})} = \sqrt{\frac{0.702 * 0.298}{2500} + \frac{0.704 * 0.296}{2500}} = 0.0129$$

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

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

We cannot conclude GPT-4 is better than Baseline! The interval contains 0.

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

Hypothesis Testing

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.

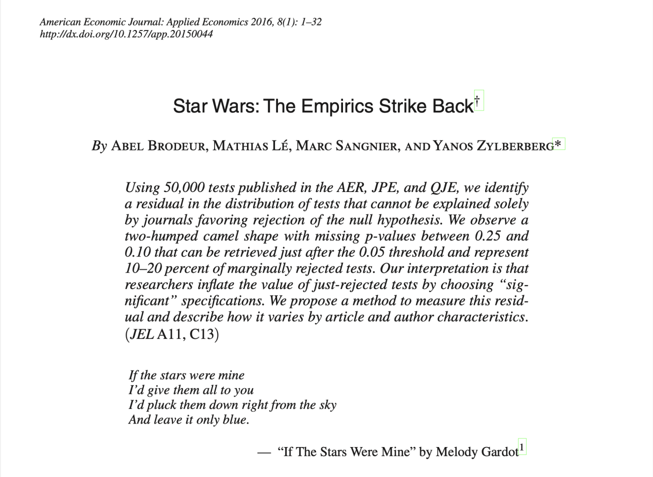
(Note: In the proportion case, this test is sometimes called a z-test)

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...
- Using the t-stat, test whether or not the Patriots are cheating in their coin tosses
- 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 LLM Eval: Gemini vs. GPT?

t-values in top Economic Journal Publications



American Economic Journal: Applied Economics 2016, 8(1): 1–32
http://dx.doi.org/10.1257/app.20150044

Star Wars: The Empirics Strike Back[†]

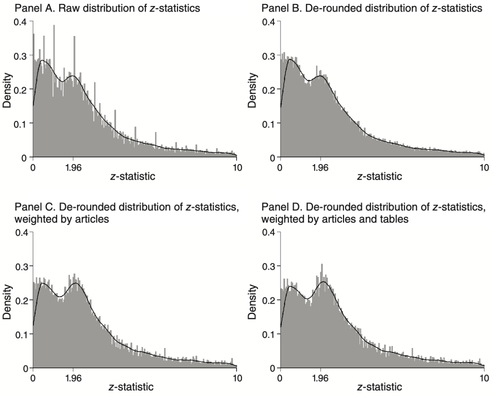
By Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg[*]

Using 50,000 tests published in the AER, JPE, and QJE, we identify a residual in the distribution of tests that cannot be explained solely by journals favoring rejection of the null hypothesis. We observe a two-humped camel shape with missing p-values between 0.25 and 0.10 that can be retrieved just after the 0.05 threshold and represent 10–20 percent of marginally rejected tests. Our interpretation is that researchers inflate the value of just-rejected tests by choosing "significant" specifications. We propose a method to measure this residual and describe how it varies by article and author characteristics. (*JEL* A11, C13)

If the stars were mine
I'd give them all to you
I'd pluck them down right from the sky
And leave it only blue.

— "If The Stars Were Mine" by Melody Gardot[1]

t-values in top Economic Journal Publications



Panel A. Raw distribution of *z*-statistics

Panel B. De-rounded distribution of *z*-statistics

Panel C. De-rounded distribution of *z*-statistics, weighted by articles

Panel D. De-rounded distribution of *z*-statistics, weighted by articles and tables

FIGURE 1. DISTRIBUTIONS OF *z*-STATISTICS

Notes: See the text for the de-rounding method. The distribution presented in the subfigure C uses the inverse of the number of tests presented in the same article to weight observations. The distribution presented in subfigure D uses the inverse of the number of tests presented in the same table (or result) multiplied by the inverse of the number of tables in the article to weight observations. Lines correspond to kernel density estimates.

Source: American Economic Review, Journal of Political Economics, and Quarterly Journal of Economics (2005–2011)

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!