Evaluating Game Experiments: A More Robust Approach for Freemium Games

Henry Phillips
Anshul Dhawan

Common Game Experiments

Common Game Experiments

- Testing performance of new features

Common Game Experiments

- Testing performance of new features
- Testing the tuning of different game variables

Common Game Experiments

- Testing performance of new features
- Testing the tuning of different game variables
- Testing new user flows

Common Game Experiments

- Testing performance of new features
- Testing the tuning of different game variables
- Testing new user flows
- Testing new, better payment flows

Step 1

Randomly split players into groups

Step 1

Randomly split players into groups

Step 2

Introduce different treatments to each group

Step 1

Randomly split players into groups

Step 2

Introduce different treatments to each group

Step 3

Evaluate performance for each group against control

Step 1

Randomly split players into groups

Step 2

Introduce different treatments to each group

Step 3

Evaluate performance for each group against control

Step 4

Pick a winner and make the change permanent

Advantages of A/B Testing

- Simple concept to understand
- Theoretically “easy to analyze”

Advantages of A/B Testing

- Simple concept to understand
- Theoretically “easy to analyze”

Hard to Implement

Standard Process

Standard Process

- Once the test runs for few days, the winner of the experiment is picked by comparing averages (e.g., Rev/User or ARPDAU) between the groups, amongst other metrics

Standard Process

- Once the test runs for few days, the winner of the experiment is picked by comparing averages (e.g., Rev/User or ARPDAU) between the groups, amongst other metrics
- A t-test is used to determine the statistical significance of results

When the Results are Positive

When the Results are Positive

Are they positive because the new feature did well ?

When the Results are Positive

Are they positive because the new feature did well ?

Or, are they positive because of the players who happen to be there ?

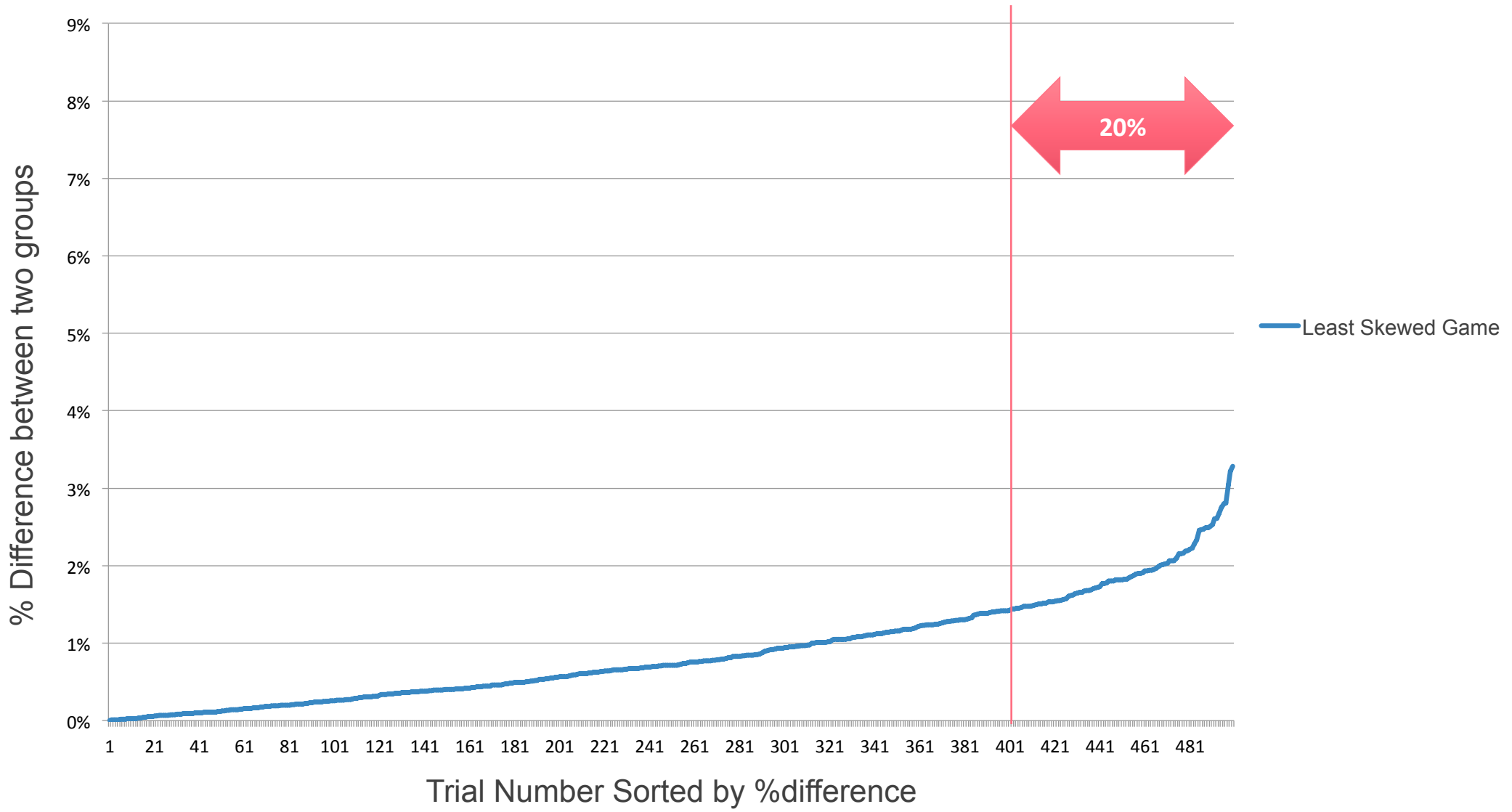
Can Signal be Overpowered by Noise Using Naïve Methods ?

We Ran 500 A/A Tests

A/A = *No difference in the experience between the two groups*

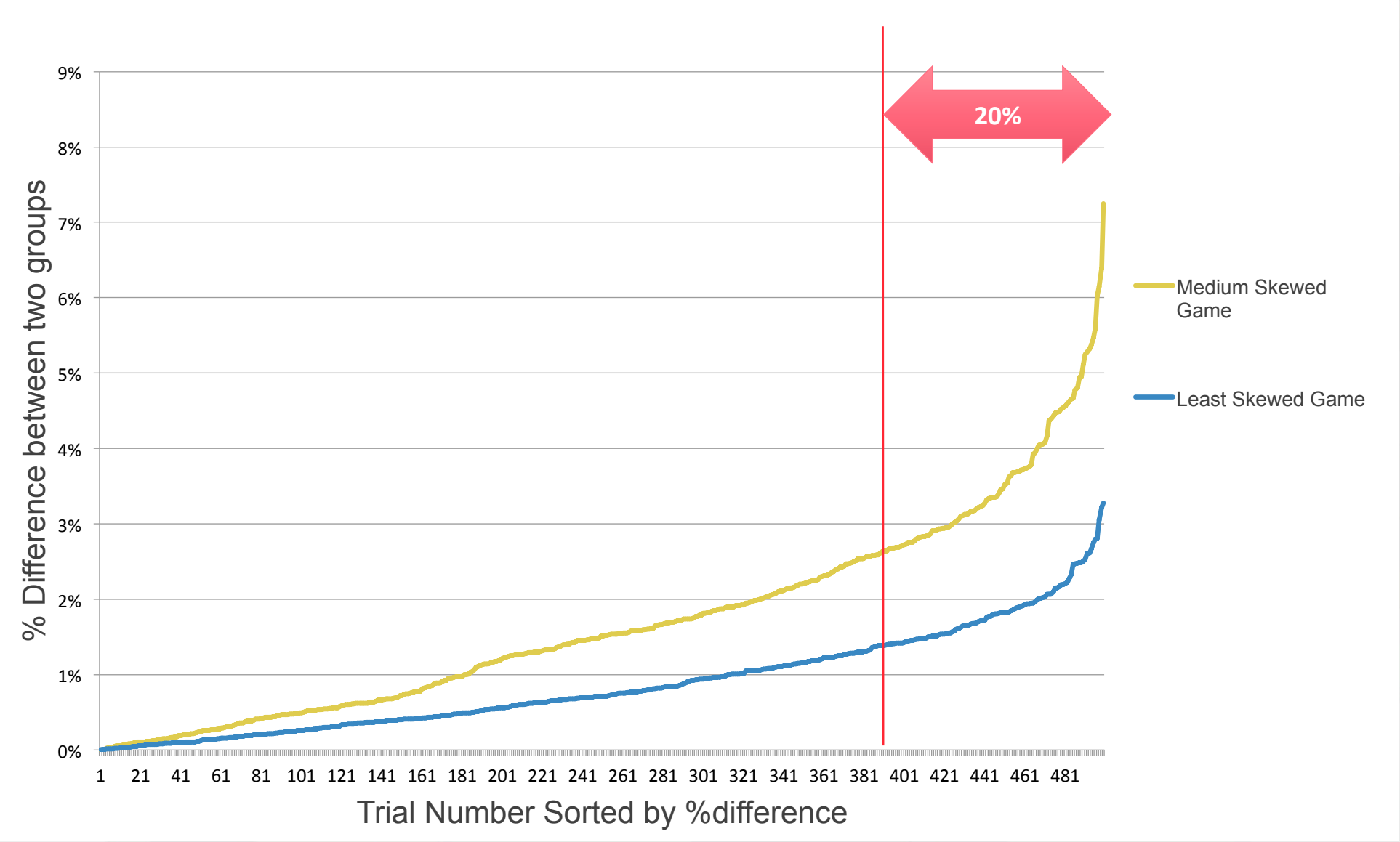
And compared the performance of the two groups

500 Random A/A Trials Comparing Rev/User

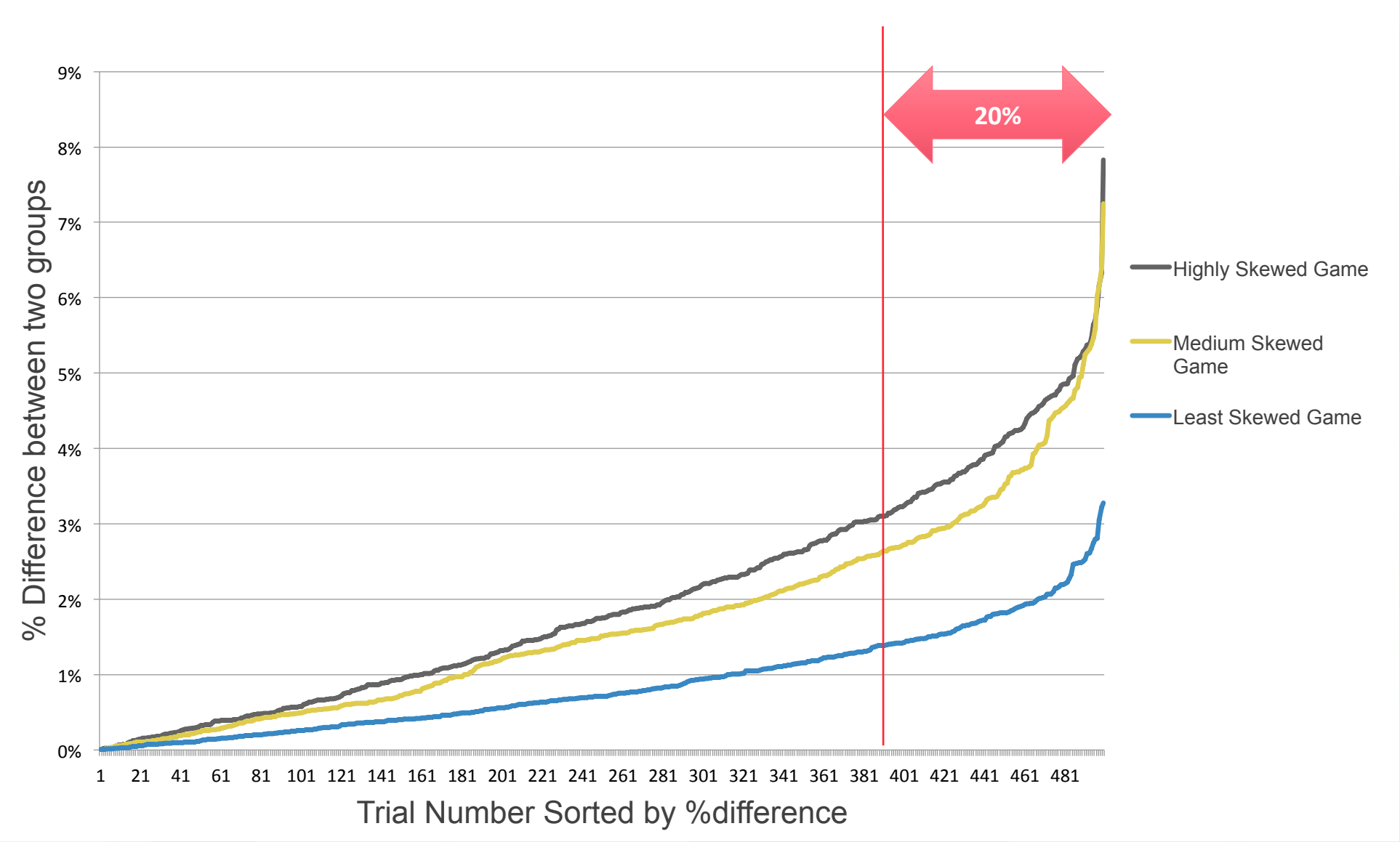


1 out of 5 times, there is a difference of > 1.4% in Rev/User between the two groups in-game that is not significantly skewed

500 Random A/A Trials Comparing Rev/User



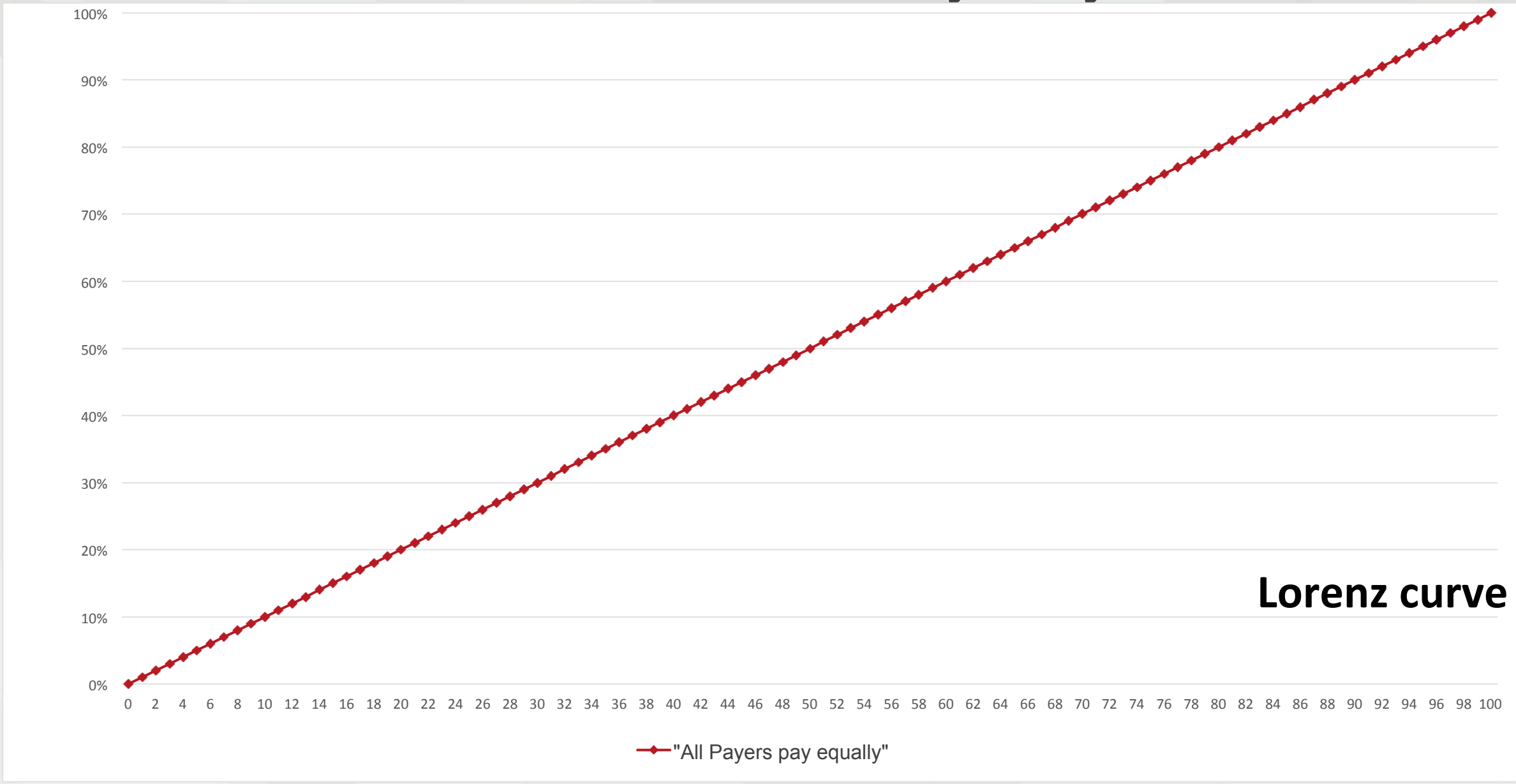
500 Random A/A Trials Comparing Rev/User



1 out of 5 times, there is a difference of > 3.2% in Rev/User between two groups in-game that is very skewed

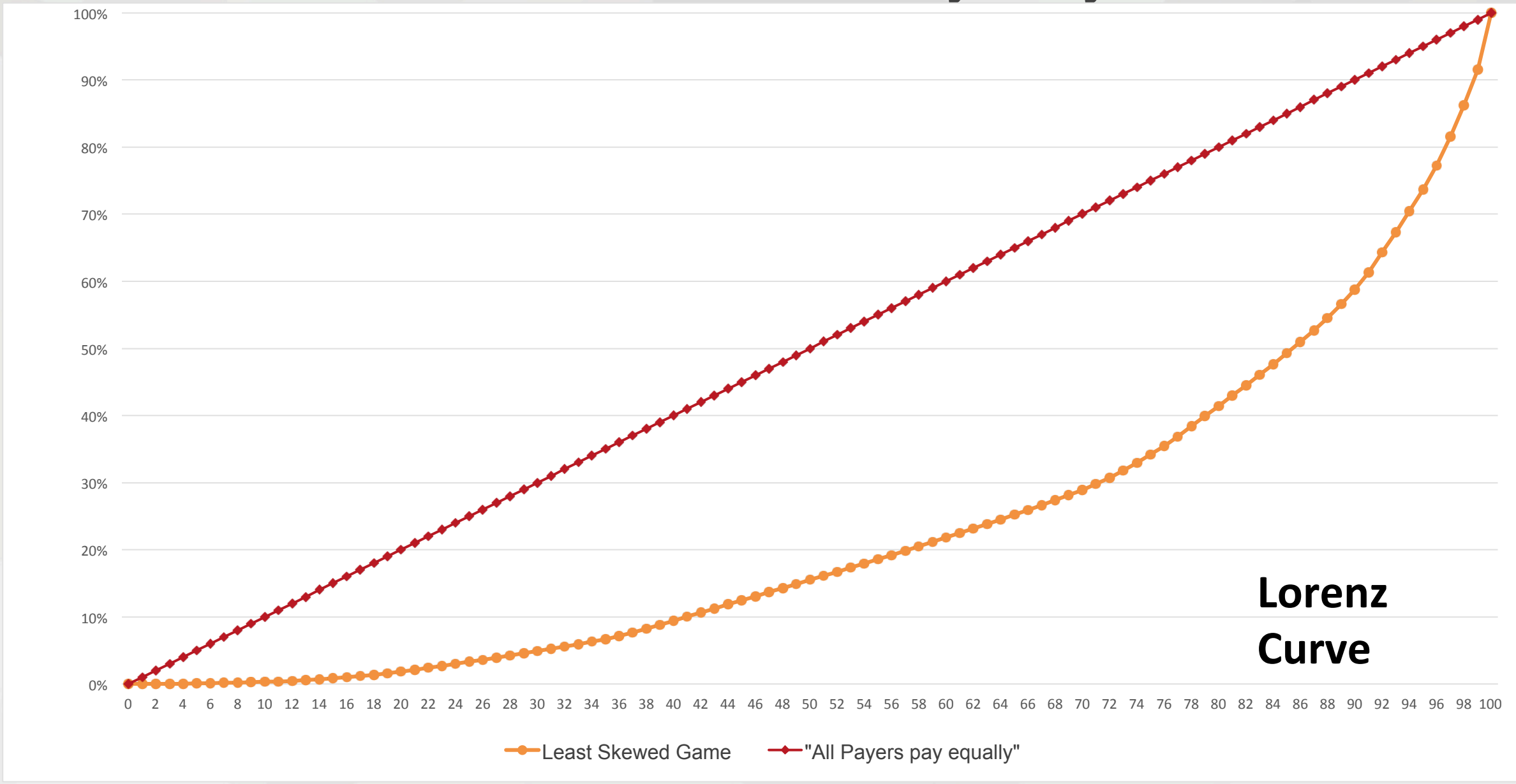
Hypothetically, lets say all payers spent the same amount of money in the game

Percent Cumulative Revenue by Payer Percentile

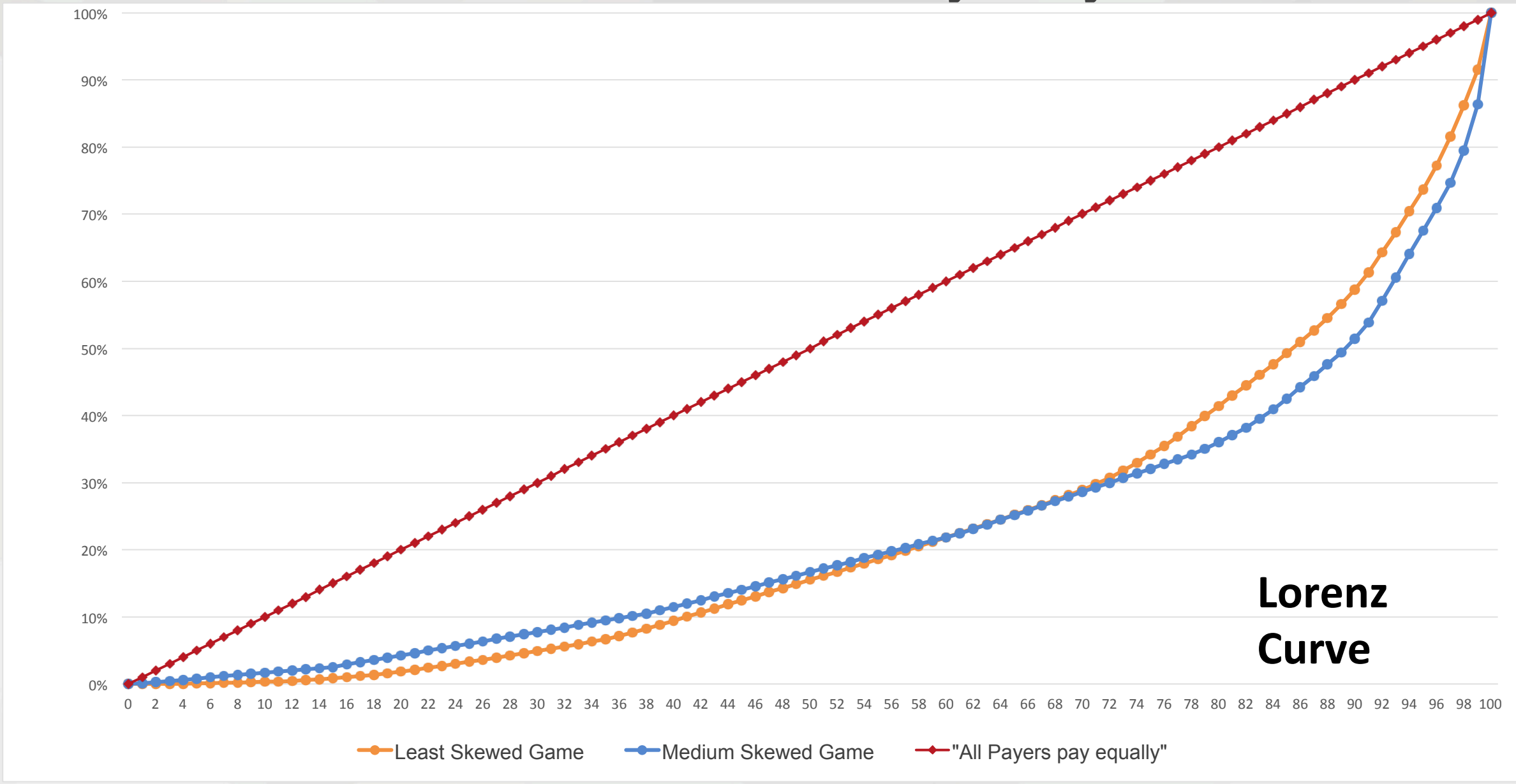


In Reality ...

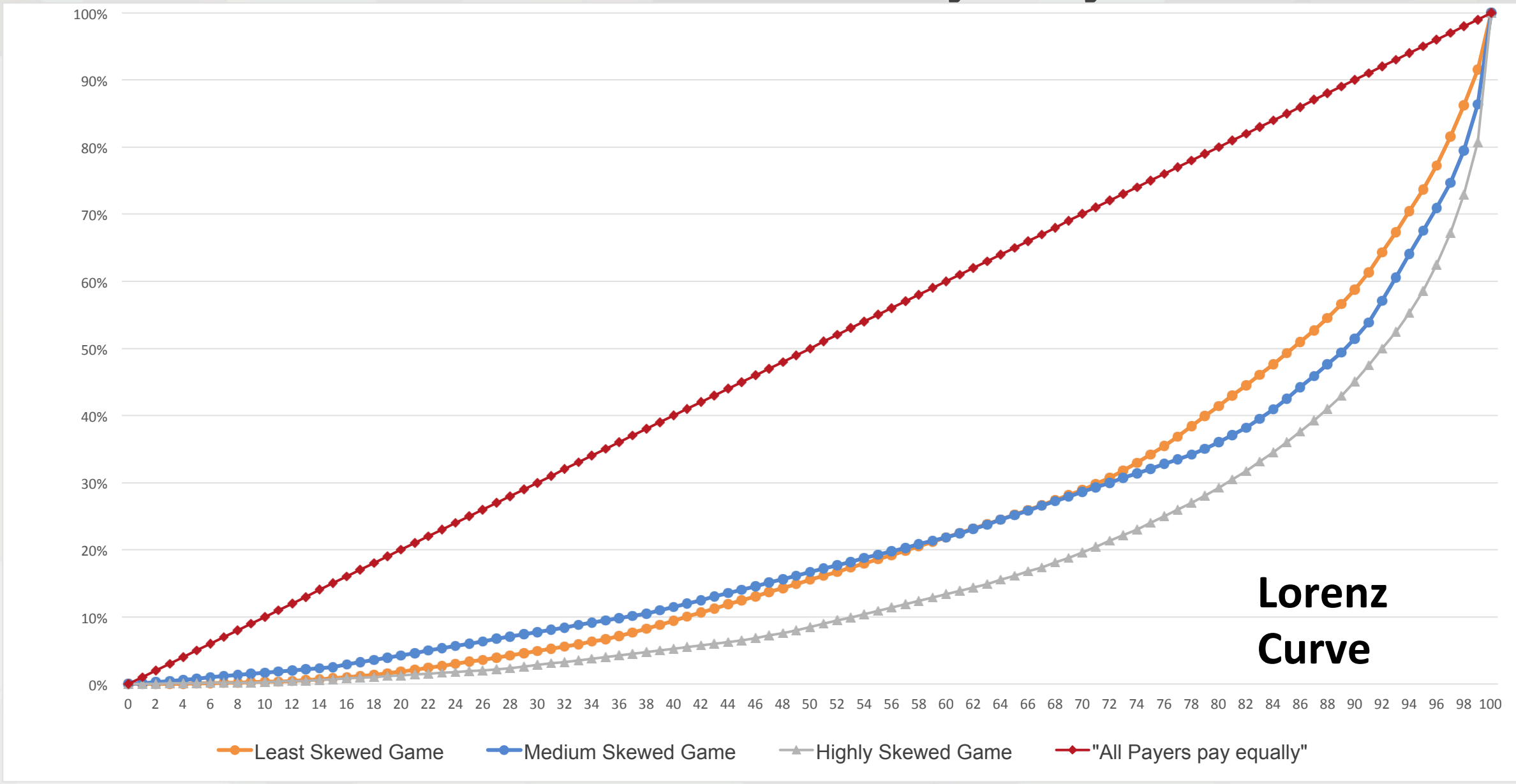
Percent Cumulative Revenue by Payer Percentile



Percent Cumulative Revenue by Payer Percentile

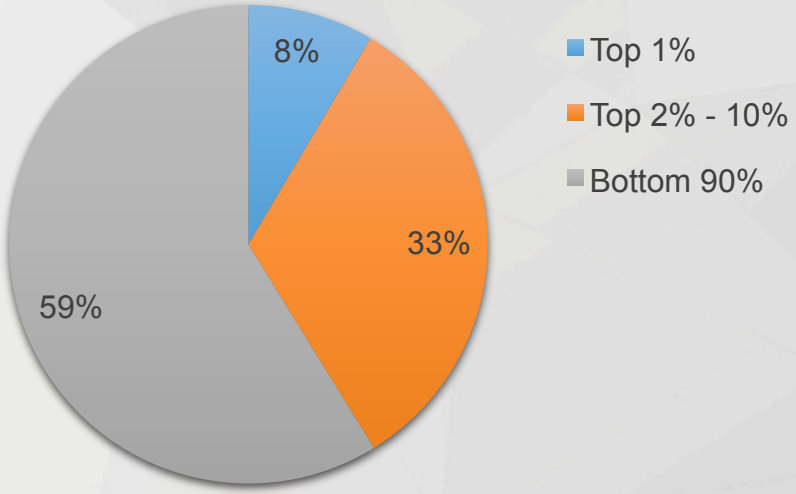


Percent Cumulative Revenue by Payer Percentile



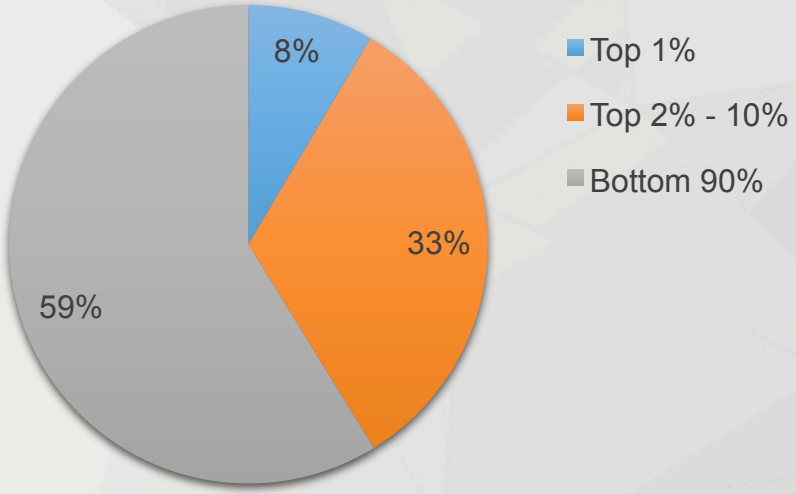
Revenue Split Between Payers

Least Skewed Game

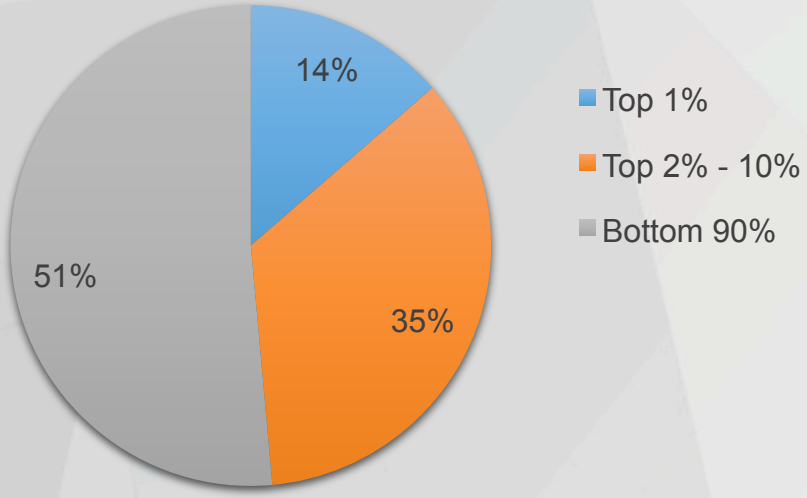


Revenue Split Between Payers

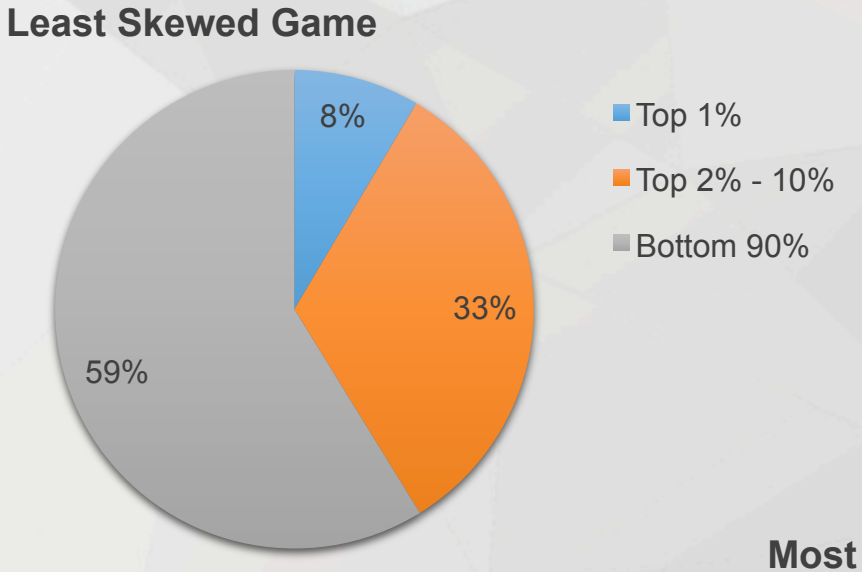
Least Skewed Game

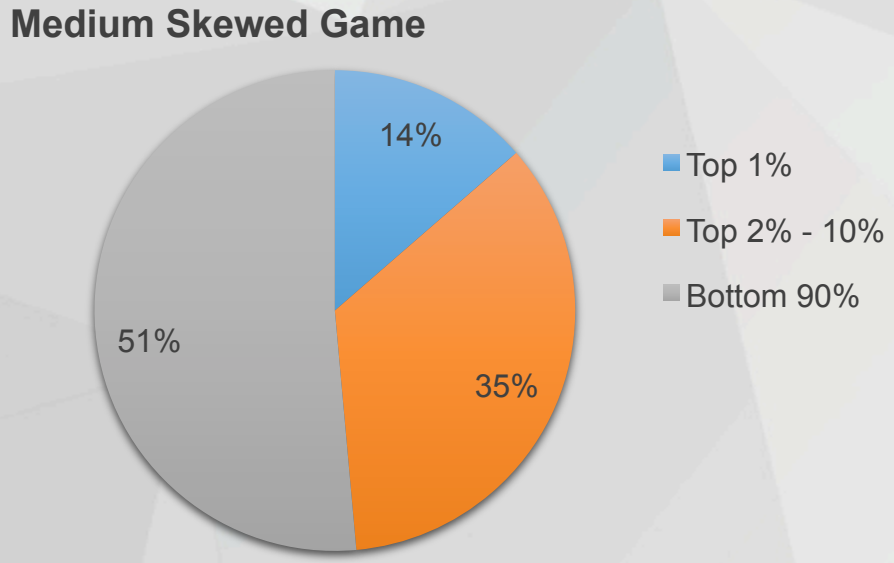


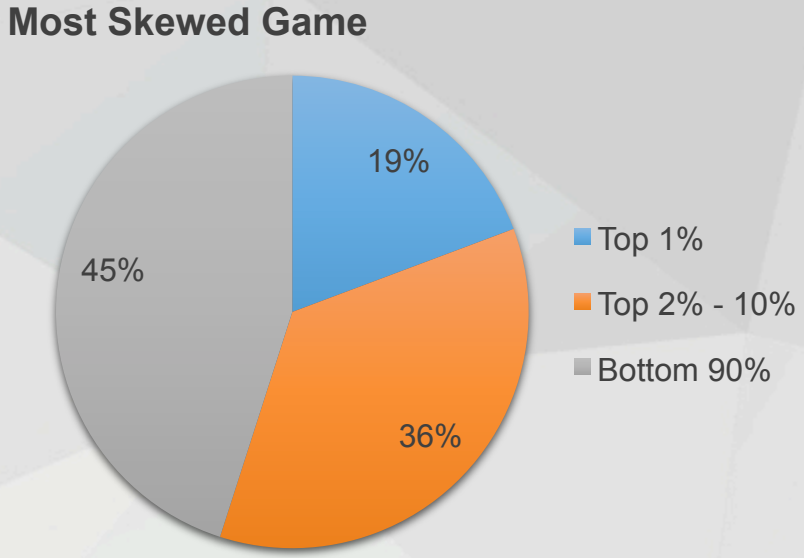
Medium Skewed Game



Revenue Split Between Payers







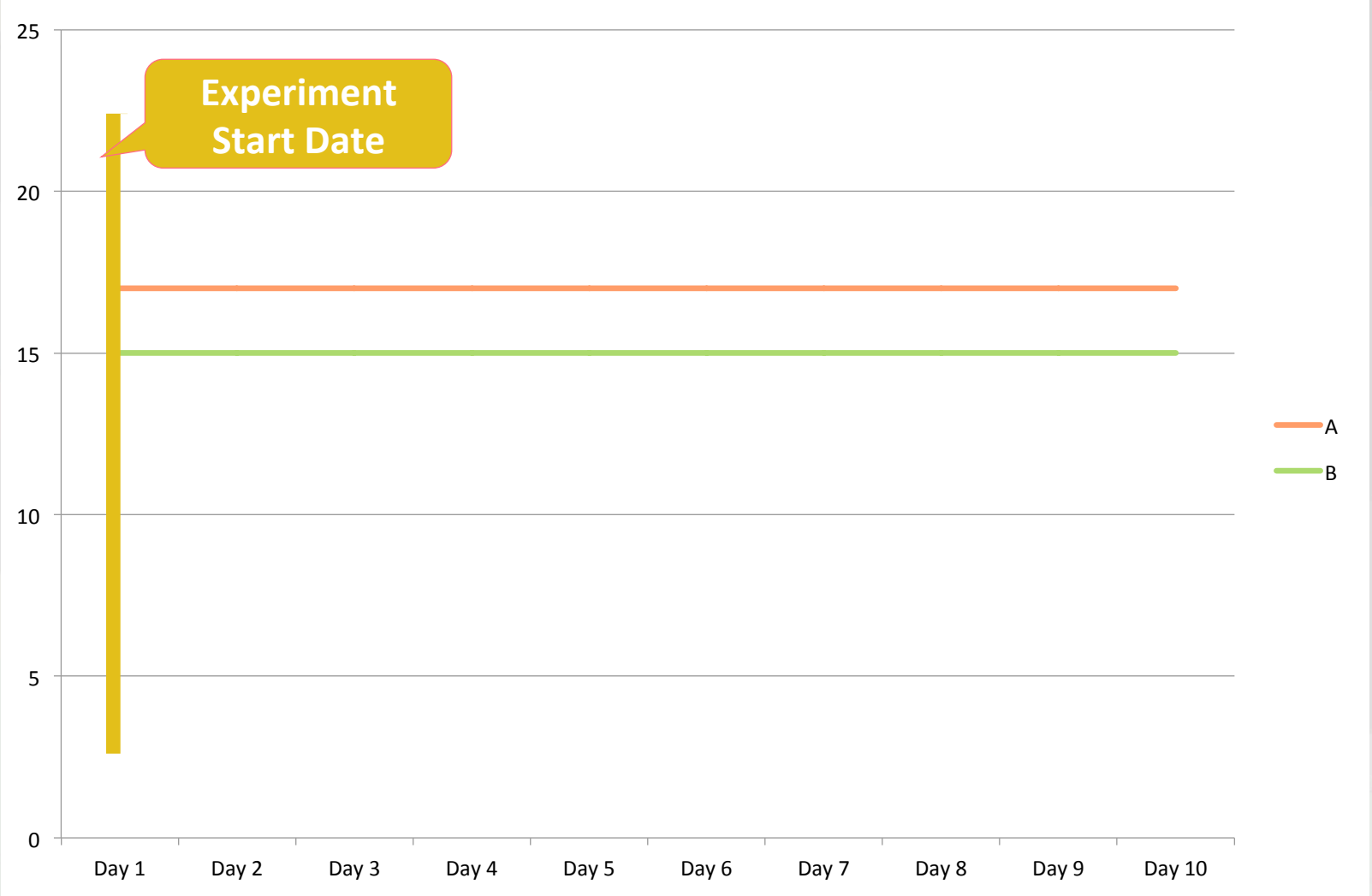
Most Games Have a Non-Normal Distribution

Most Games Have a Non-Normal Distribution

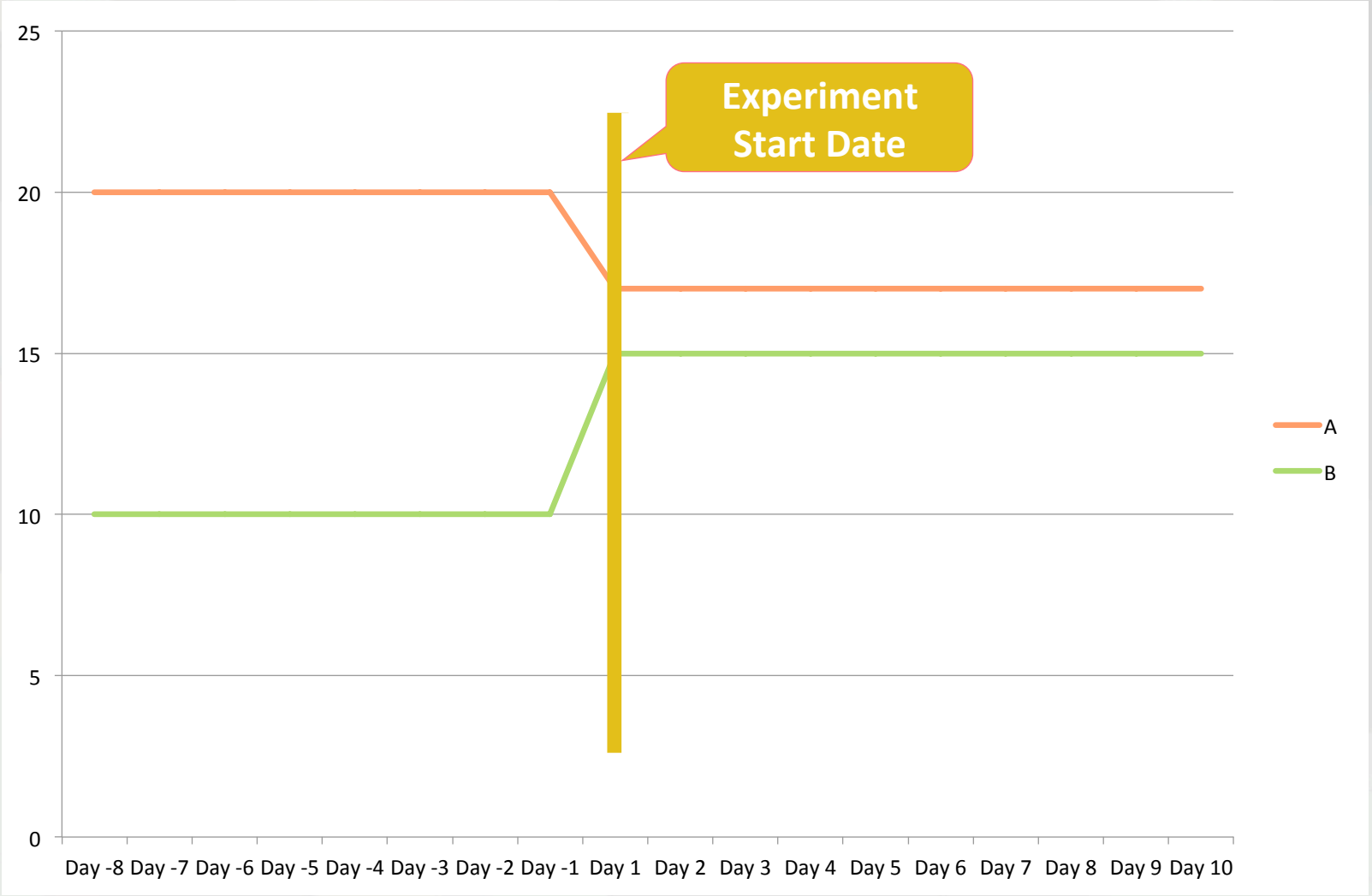
Unequal Split of Top Spenders Can Cause Bias in the Split of Users

Results that look great, might in reality be underperforming variations

Variant A Might Look to be Performing "Better" Than B



But in Reality it Might be Performing Worse



Goal

Fast decision making that allows rapid iteration

Hide the complexity of understanding distributions

Empower product managers to run more experiments

Challenge

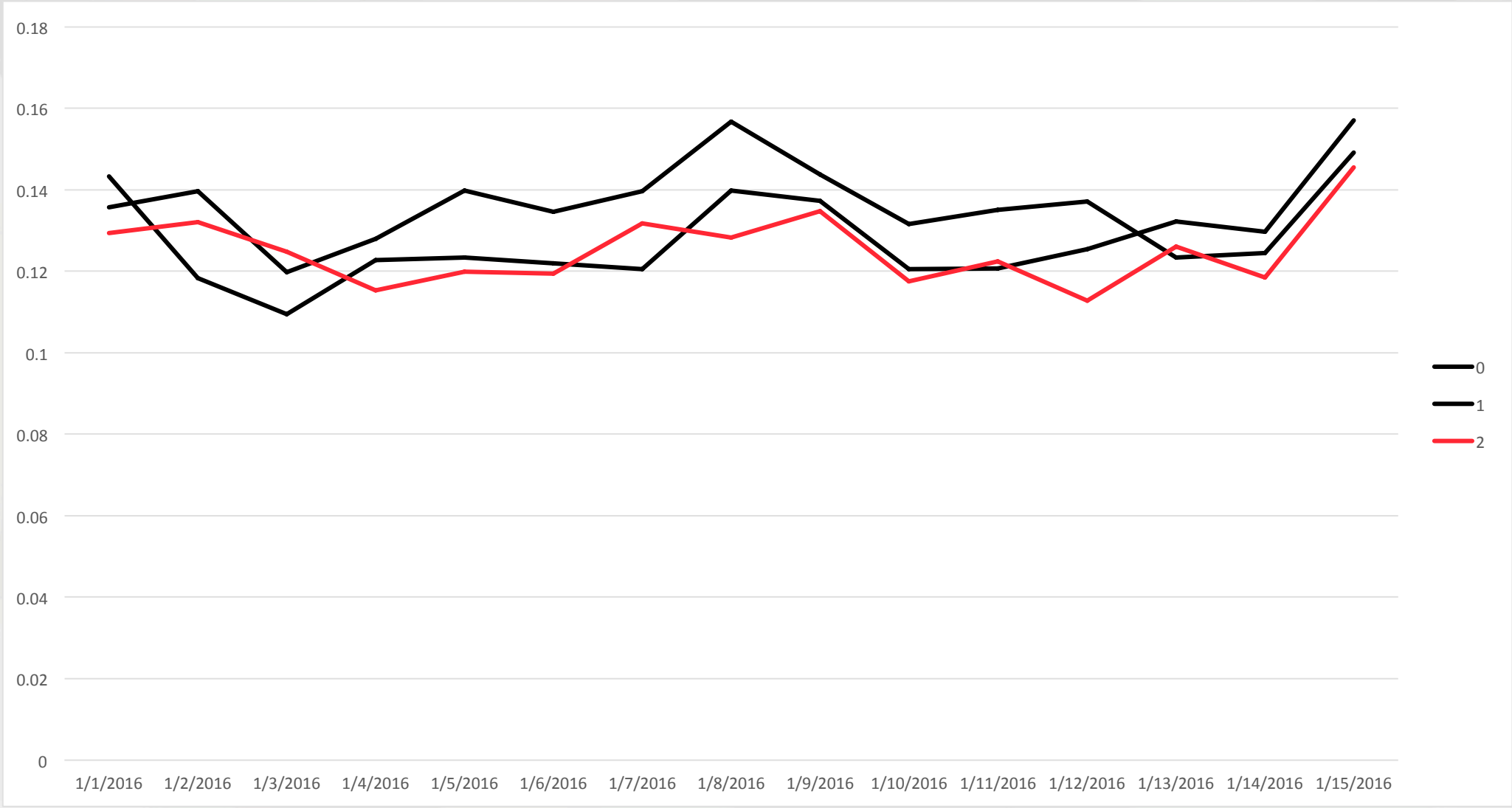
Data Scientists can’t analyze every experiment

Decision makers don’t have infinite time to make a decision

Methodology	
Naive	Highly Prone to Top Spender Imbalance

Methodology	
Naive	Highly Prone to Top Spender Imbalance
Dual Control	

Dual Control



Methodology	
Naive	Highly Prone to Top Spender Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls

Methodology	
Naive	Highly Prone to Top Spender Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls
Mann-Whitney U	

Mann Whitney U

Rank	Revenue	Variant
1	$170.0	B
2	$133.0	A
3	$129.0	A
4	$110.0	A
5	$90.0	B
6	$88.0	B
7	$75.0	A
8	$66.0	A
9	$65.0	A
10	$60.0	B
11	$59.0	B
12	$58.0	B
13	$55.0	B
14	$50.0	A
15	$48.0	A
16	$46.0	B

Sum of B Rank
74
Sum of A Rank
62

Methodology	
Naive	Highly Prone to Top Spender Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls
Mann-Whitney U	Statistical Sledgehammer That Gives Perfect Results but Doesn't Tell the Magnitude of the Change

Methodology	
Naive	Highly Prone to Top Spender Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls
Mann-Whitney U	Statistical Sledgehammer That Gives Perfect Results but Doesn't Tell the Magnitude of the Change
Pre-Post	

Pre - Post

Compare the difference between the performance of the a group of users before and after the test i.e sum(pre-test values)/count(users) to sum(post-test values)/count(users)

Pre - Post

	Pre-Test Values	Post-Test Values	Difference
Control	X1	Y1	Z1 = (Y1-X1)/ X1
Test	X2	Y2	Z2 = (Y2-X2)/ X2

Pre - Post

	Average	10th	25th	median	75th	90th
Normal	1.15%	0.18%	0.44%	0.95%	1.63%	2.44%
Pre-Post	0.86%	0.15%	0.34%	0.73%	1.26%	1.77%
% Gain	-25.01%	-15.99%	-21.74%	-22.40%	-22.71%	-27.54%

- Pre-post reduced the average noise to 0.86% (25.01% less)

Methodology	
Naive	Highly Prone to Whale Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls
Mann-Whitney U	Statistical Sledgehammer that Gives Perfect Results but Doesn't Tell the Magnitude of the Change
Pre-Post	Doesn't Account for Non-Payers and New Installs

Methodology	
Naive	Highly Prone to Whale Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls
Mann-Whitney U	Statistical Sledgehammer that Gives Perfect Results but Doesn't Tell the Magnitude of the Change
Pre-Post	Doesn't Account for Non-Payers and New Installs
Neighborhood Band Normalization	

Neighborhood Band Normalization

Performance of the Variant =
Sum of Actual Results/
Sum of Estimated Results

Neighborhood Band Normalization

User Rank	Pre Rev	Post Rev	Estimation	Variant
…				
101	7	2.86		A
102	6	0.31		A
103	5	2.35		B
104	4	1.74	1.59563435	A
105	3	1.96		B
106	2	1.83		B
107	1	0.12		A
…				

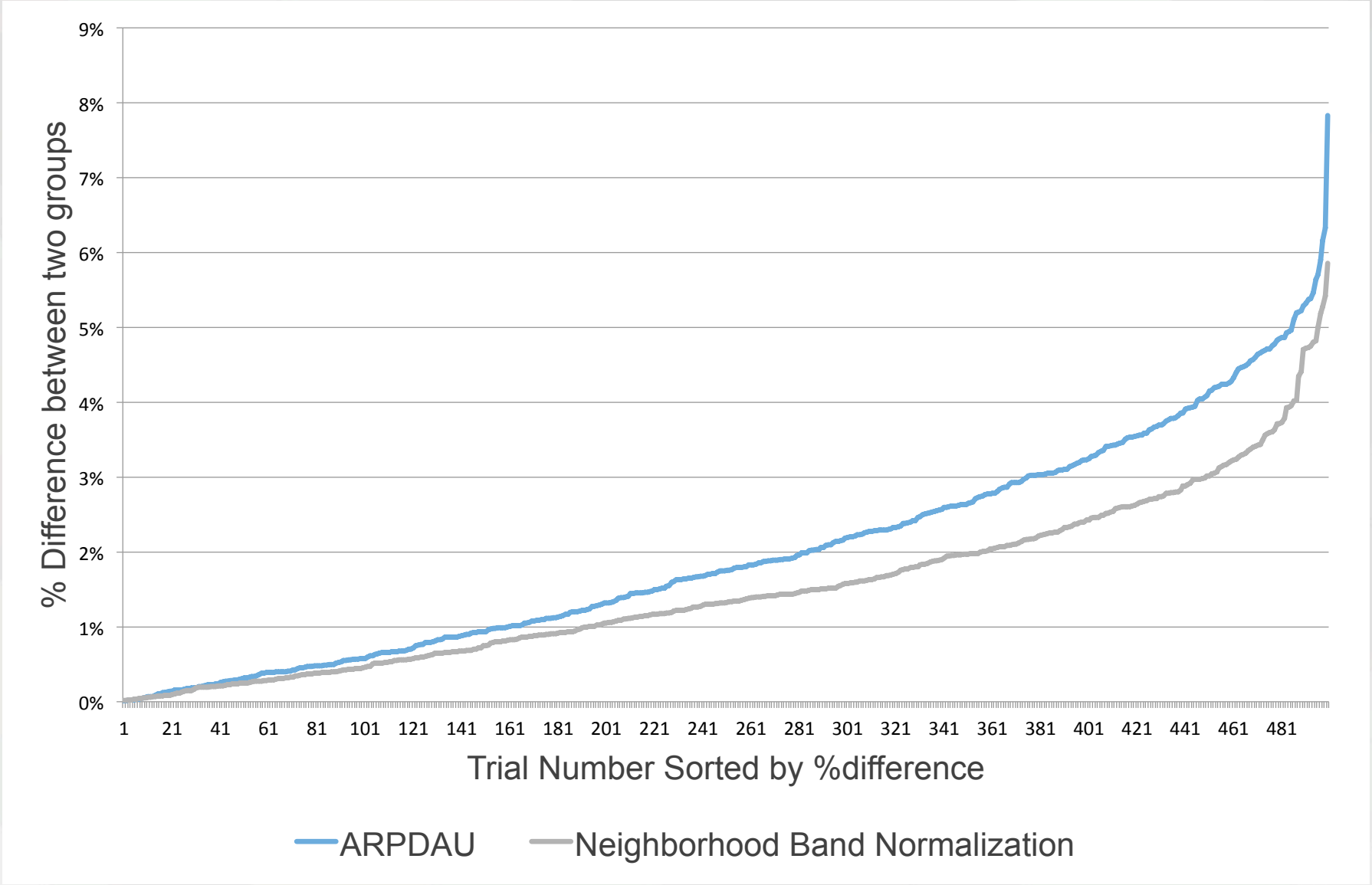
Neighborhood Band Normalization

- Rank users bases on prior-to-test features
  - Prior rev, prior game actions, prior engagement, geo
- Post Rev estimation = Average post rev of the 100 users ranked above and below them (w/ adjustment factors for those who don't have 100 above or 100 below)
- Makes estimations for non-payers and installs

Neighborhood Band Normalization

Noise	Avg	10th pctile	25th pctile	Median	75th pctile	90th pctile
Standard	1.15%	0.18%	0.44%	0.95%	1.63%	2.44%
Pre-post	0.86%	0.15%	0.34%	0.73%	1.26%	1.77%
Band	0.79%	0.13%	0.29%	0.65%	1.14%	1.65%
Pre vs Std	-25.01%	-15.99%	-21.74%	-22.40%	-22.71%	-27.54%
Band vs Std	-31.43%	-28.33%	-34.28%	-30.74%	-29.86%	-32.36%

500 Random A/A Trials for Highly Skewed Games



Normalizing Actual Results by Predictive Results Reduces the Noise by 31%

It is Important to Take Prior Information Into Account

3 variations of a feature that grants different rewards showed the following result based on Rev/User

3 variations of a feature that grants different rewards showed the following result based on Rev/User

	% Difference from Control	
	Naïve	Neighborhood Band Normalization
Variation 1	7.33%	
Variation 2	4.62%	
Variation 3	2.23%	

The group that was exposed to Variation 1 , had 40% more Top 1% payers than Control

3 variations of a feature that grants different rewards showed the following result based on Rev/User

	% Difference from Control	
	Naïve	Neighborhood Band Normalization
Variation 1	7.33%	-7.24%
Variation 2	4.62%	-7.17%
Variation 3	2.23%	11.65%

Methodology	
Naive	Highly Prone to Whale Imbalance
Dual Control	Only Provides a Visual Representation of Natural Variance Between Controls
Mann-Whitney U	Statistical Sledgehammer that Gives Perfect Results but Doesn't Tell the Magnitude of the Change
Pre-Post	Doesn't Account for Non-Payers and New Installs
Neighborhood Band Normalization	It's Better on Average But Not Always

Always Room for a Better Methodology

Always Room for a Better Methodology

Continued investment into Bayesian Methods
to find a more robust approach that can withstand any distribution found in games

Common Pitfalls

- Setting up experiment correctly
- Testing things that are actually meaningful
- Too many experiments going on
- Not analyzing the right metrics
- Not understanding how your top payers behave

Games Are Pieces of Art as Much as Science

- Sometime testing is only good to get a directional sense
- Don’t let data govern you
- Trust your intuition
- Common sense over data

Thank You

Questions ?

Henry Phillips : hphillips@zynga.com

Anshul Dhawan: adhawan@zynga.com | @theanshuldhawan