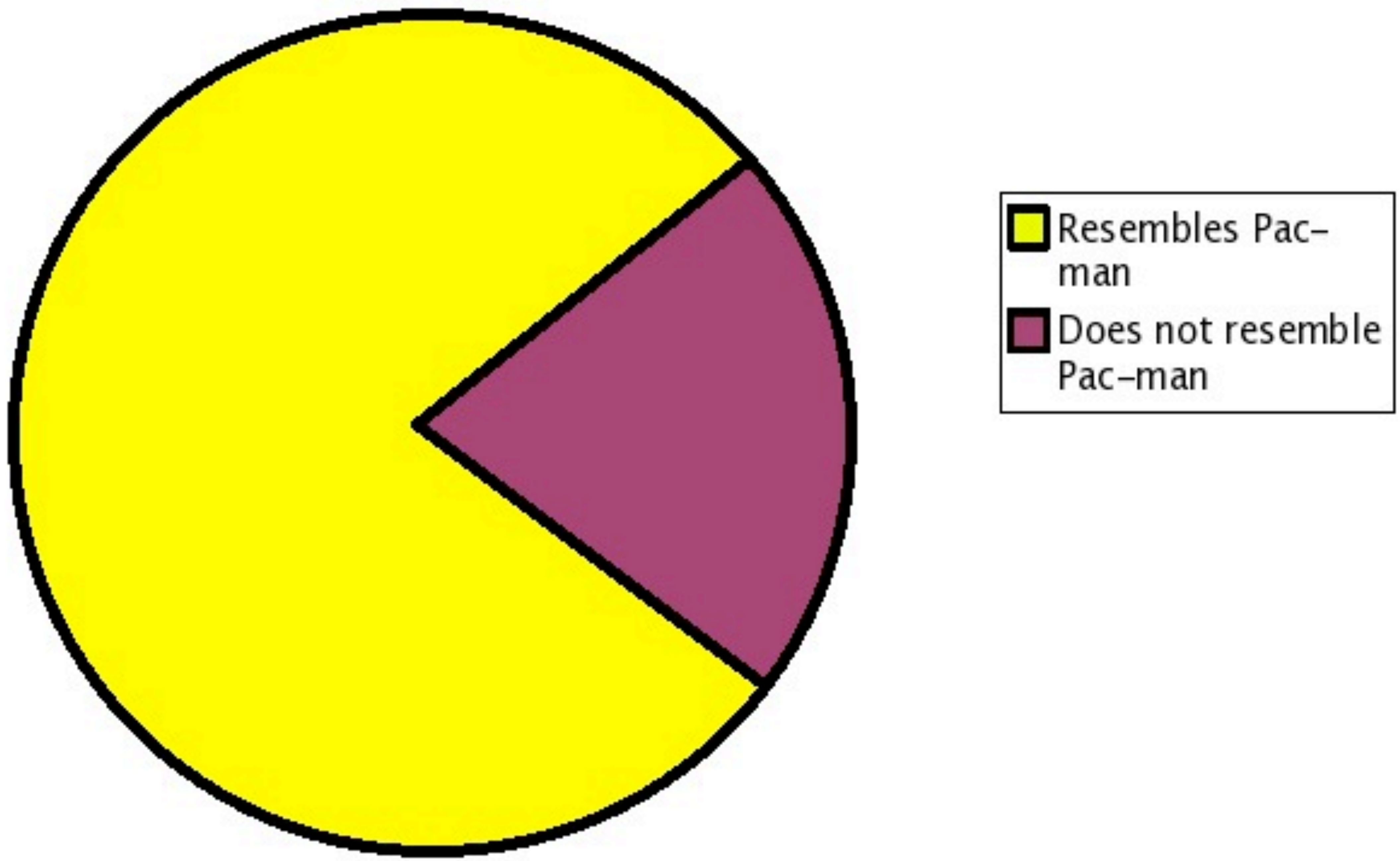
Percentage of Chart Which Resembles Pac-man



ANALYZING EXPERIMENTS
COMPARING RATES

Scott Klemmer

Analyzing your data in 3 questions

1. **What does my data look like?**
   Explore your data graphically
   Plot all your data
   Plot several different summaries
2. **What are the overall numbers?**
   Aggregate statistics for each condition
   Usually mean and standard deviation
3. **Are the differences "real"?**
   Compute significance (p value)
   Likelihood that results are due to chance

Say I have a coin

What attributes does our statistic need?

Pearson's Chi-Squared Test

Karl Pearson, 1900, http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

‘Normal’ outcome variance

The Null Hypothesis

Critical Values for Chi-Squared

df \area	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	4E-05	2E-04	0.001	0.004	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75

table from http://www.statsoft.com/textbook/distribution-tables/

Example: Is this a balanced coin?

- 20 tosses. 13 heads. At $p<0.05$, can we reject the null hypothesis that there is no difference between the test coin and an unbiased coin?

Example: Is this a balanced coin?

df \area	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	4E-05	2E-04	0.001	0.004	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75

table from http://www.statsoft.com/textbook/distribution-tables/

What if the trend continued?

- Say we tossed a coin 60 times, and saw the same pattern: 39 heads out of 60

- We can reject the null hypothesis with 98% confidence
- Note (if the trend is robust) increasing sample size by a factor of 3 decreases the probability of a false positive by a factor of 9

df \area	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.75**	**0.5**	**0.25**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	4E-05	2E-04	0.001	0.004	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75

table from http://www.statsoft.com/textbook/distribution-tables/

Example: Improved click-throughs?

- A web site has a button labeled “sign up”. 10% of visitors click the button.
- To try and improve traffic, they change the button to “learn more”, and start gathering data.
- Over a week, there were 1000 visitors to the site. 119 clicked the “learn more” button.
- Can we say with confidence that the “learn more” button has a higher click-through rate than the “sign up” button?

Example: Improved click-throughs?

- df=1
- The odds that the observed difference happened by chance is (just barely) $p<0.05$
- The change (probably) improved click rate

df \area	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	4E-05	2E-04	0.001	0.004	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75

Statistical testing

- Formalizes "we're pretty sure"
- Helps you generalize (or not) from small samples

This insight owes a lot to beer

Image: http://en.wikipedia.org/wiki/File:St._James%27s_Gate_Brewery,_Dublin,_Ireland.jpg
Story: http://en.wikipedia.org/wiki/Student's_t-test

For ‘normal’, *continuous* data

- T-tests (compare 2 conditions)
- ANOVA (compare >2 conditions)

Data Often Ain’t ‘Normal’

Handling non-'normal' data

- Knowing is half the battle
- Run A/A tests
- Use randomized testing

Summary

- To get a feel for your data, graph it all
- Statistics provides tools to distinguish 'real' trends from 'mirages'
- We learned a common technique for comparing rates: the chi-squared test

To Learn More...

- *Practical Statistics for HCI*, Jacob Wobbrock, http://depts.washington.edu/aimgroup/proj/ps4hci
- *Doing Psychology Experiments*, David W. Martin
- *Statistics as Principled Argument*, Robert P. Abelson
- *Learning to use statistical tests in psychology*, Judith Greene, Manuela D'Oliveira