Supplemental Resource: Brain and Cognitive Sciences
Statistics & Visualization for Data Analysis & Inference
January (IAP) 2009

Statistics and Visualization for Data Analysis: Resampling etc.

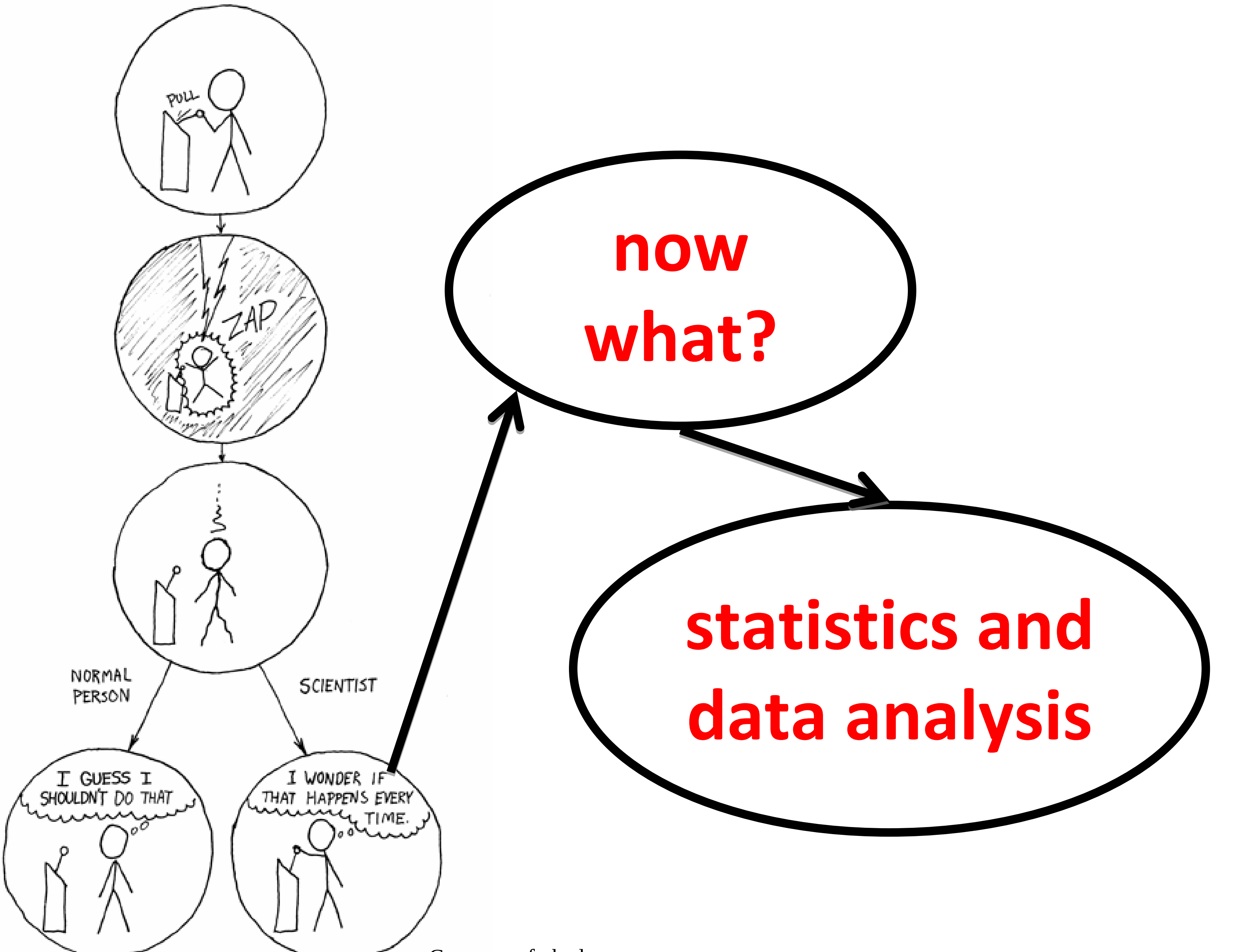
Mike Frank & Ed Vul

IAP 2009

Today’s goal

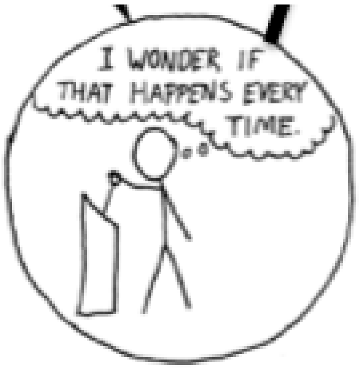
- Classically, statistics is full of equations.
- This is (partly) because computers have not been around for long

- **Convey the principles behind frequentist statistics using only numerical methods (i.e., by using the brute force of computers)**

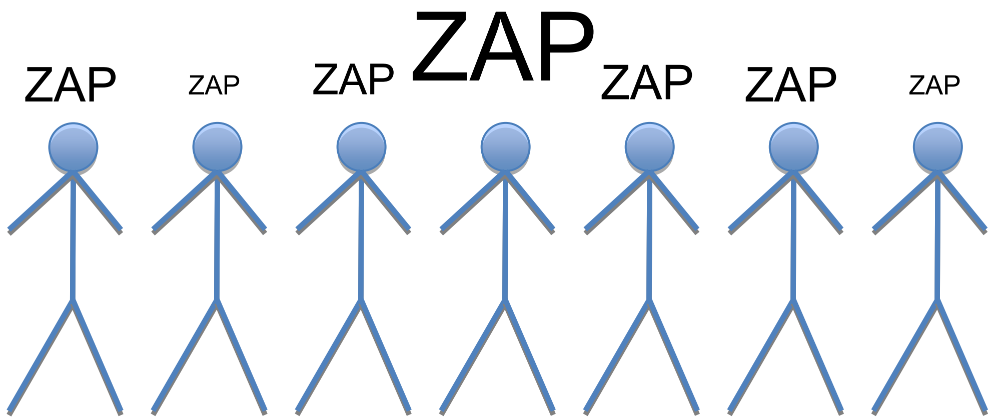


Courtesy of xkcd.org

Does that happen every time?

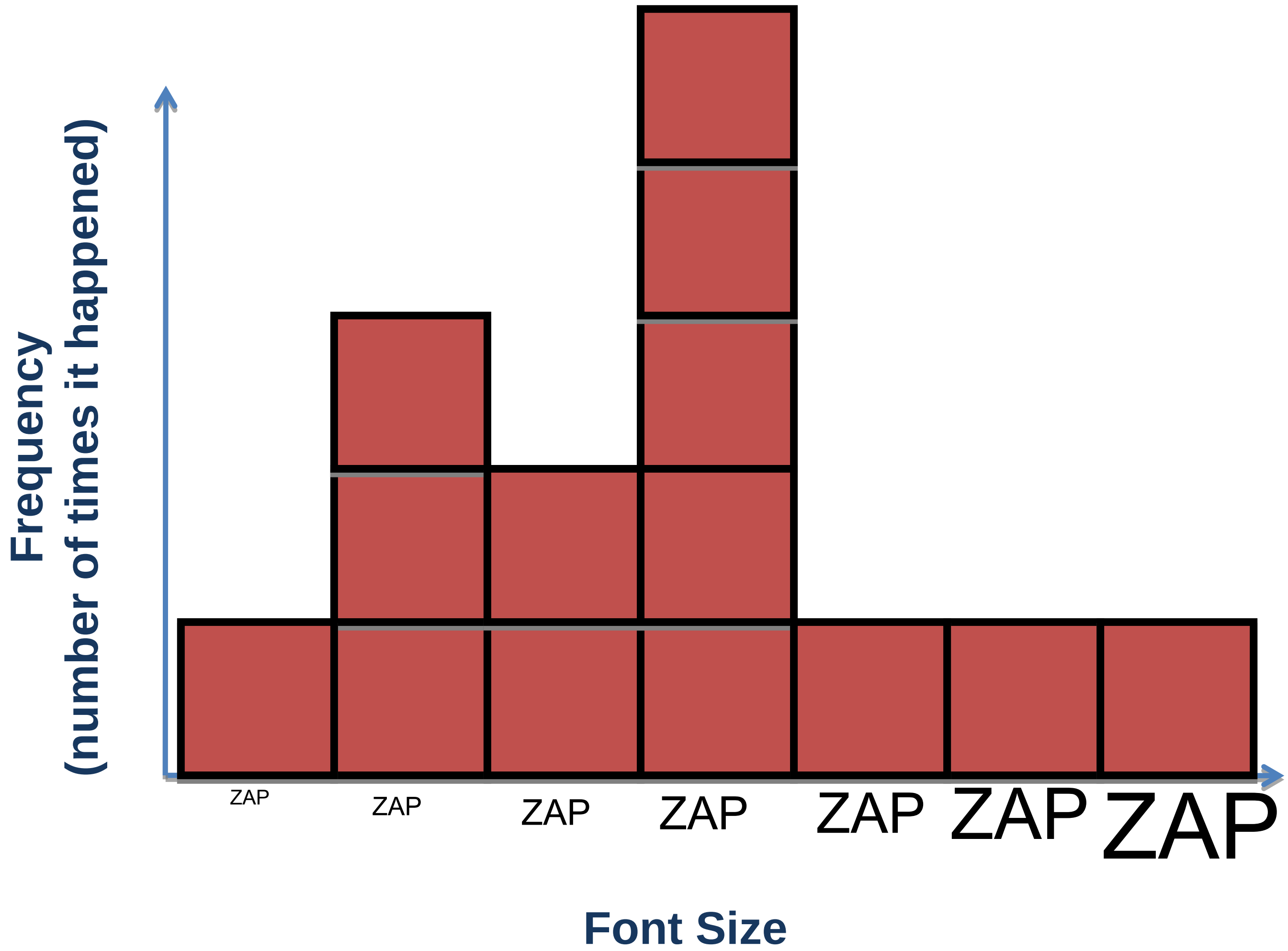


Courtesy of xkcd.org



now what?

What has happened?



What’s going to happen next?

- We don’t know.
- Let’s assume ‘more of the same’.
- ‘More of the same’:
  - Some process was producing events.
  - Events were
    - Independent
    - Identically distributed
- Assume “more independent, identically distributed events will follow”

Uncertainty

- We don’t know exactly what will happen if we touch the podium again.
- However, we have some data.
- The data allow us to make predictions.
- We can measure our uncertainty about what will happen with probability.

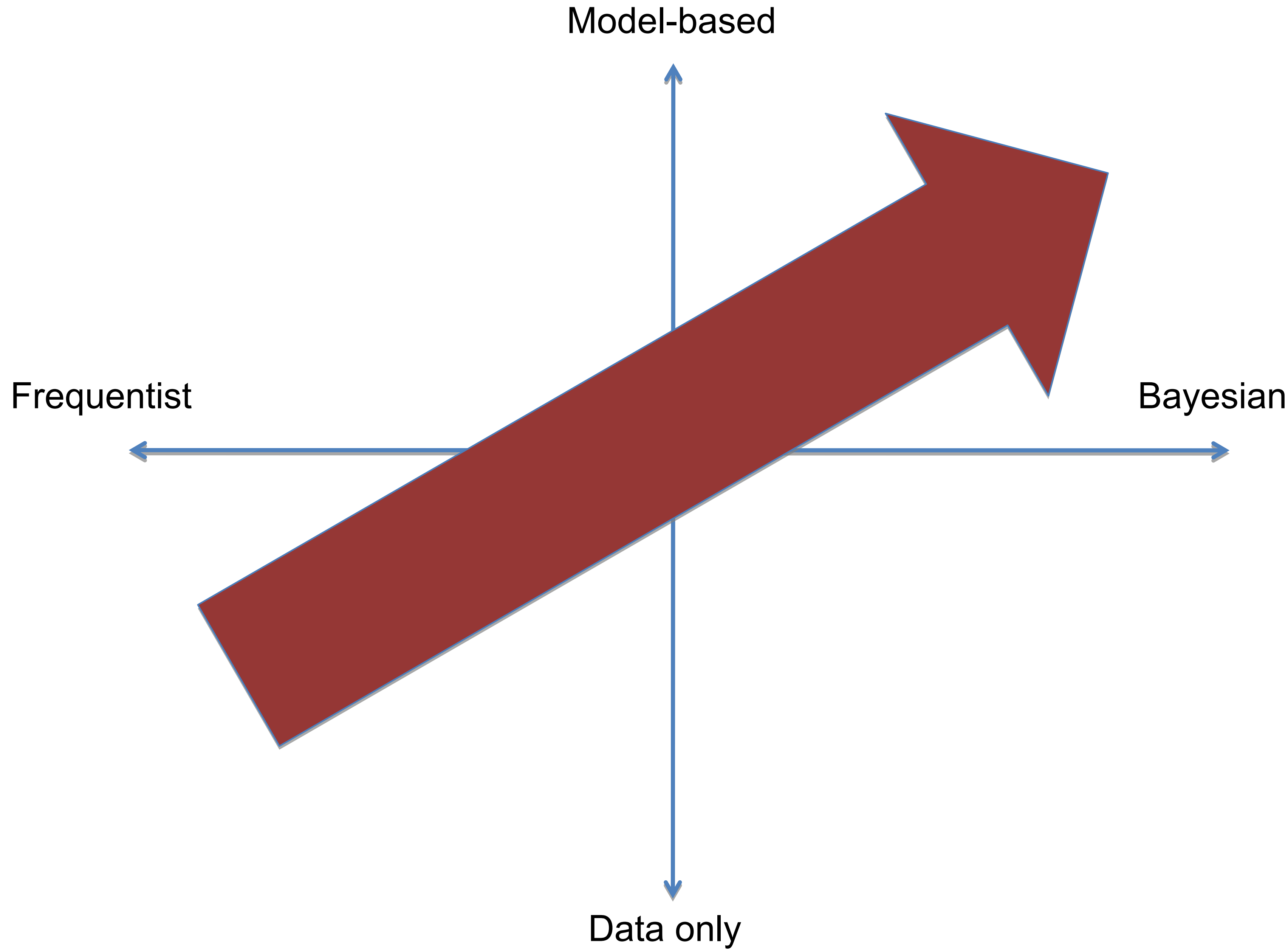
“Probability”?

- Frequentist:
  One specific event will happen next.
  Another specific event will happen after that.
  All we can say is that over many such events, the frequency of a specific one occurring will match the frequency we observed up to now.
  - Probability is long-run frequency.
- Bayesian:
  I don’t know what will happen next, but I have some *beliefs* about what it could be. These beliefs follow the laws of probability. (My beliefs will reflect more than just the data.)
  - Probability is degree of belief.

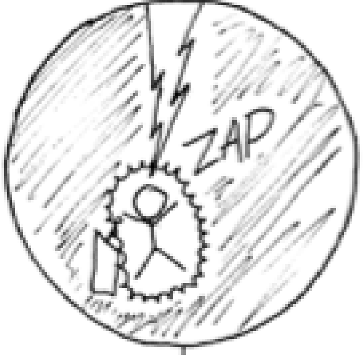
Frequentist or Bayesian?

- Most statistics you have been exposed to are ‘frequentist’.
  - Interpretations of e.g., ‘confidence intervals’ are rather weird.
  - Prior beliefs (such as theory, or good reason) don’t matter.
- We will be frequentist for most of today, but there are reasonable Bayesian interpretations of what we are doing.
- Let’s not worry about it for now.

Our class trajectory



What will happen next?



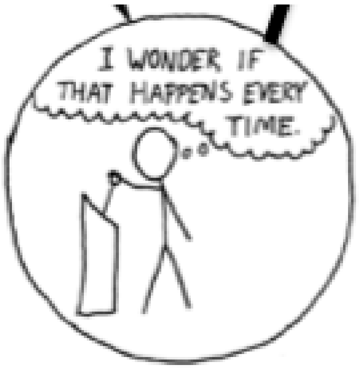
Courtesy of xkcd.org

- One of the prior events will repeat with a probability matching its previous frequency.

- So… we can just draw samples (with replacement!) from the previous data to predict future data.

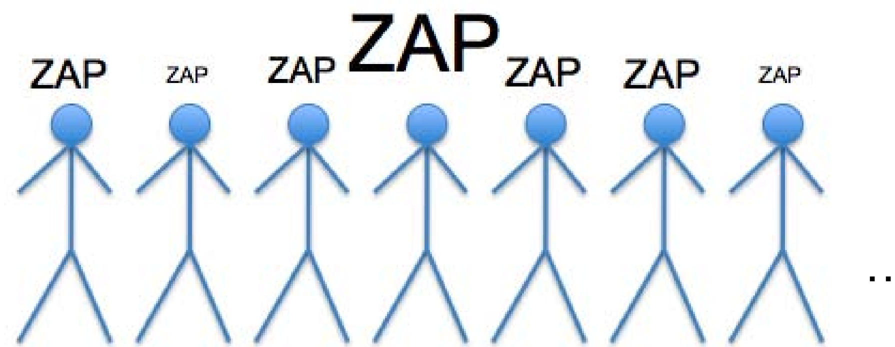
- This is **resampling**

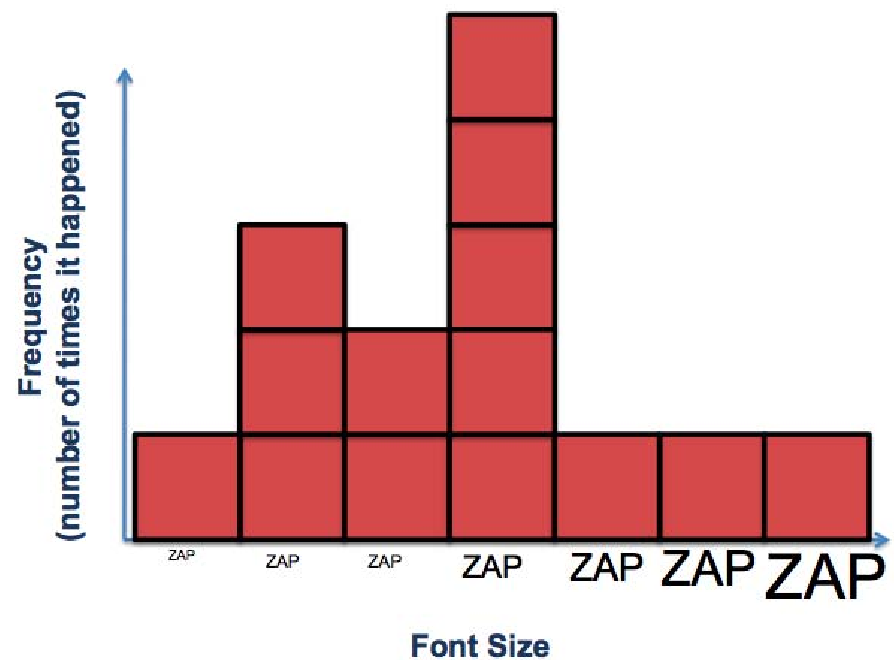
It’s not that simple



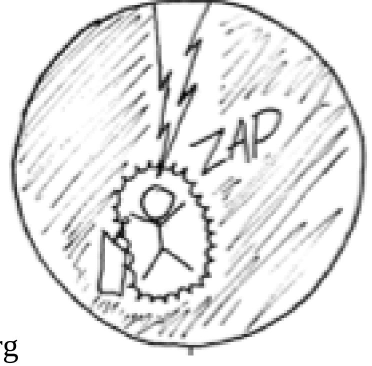
Courtesy of xkcd.org

now what?





What do we want to know?



Courtesy of xkcd.org

- The mean font size of a zap?
- Do zaps happen more often in this case than otherwise?
- How much bigger are average font sizes at the podium?
- If we got zapped at the podium or somewhere else, which zap would have a bigger font size?

The mean font size of a zap?

ZAP

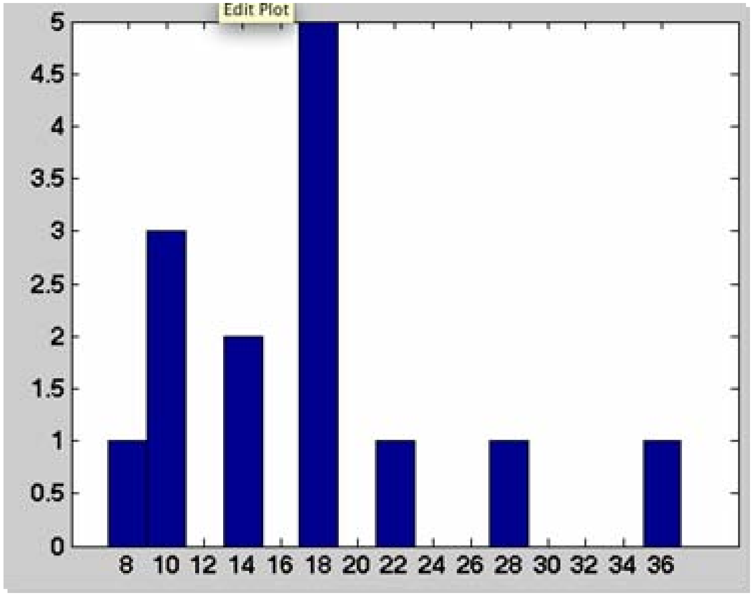
- Great. Wait. We’re not done.

- What we really want to be able to do is predict the average font size of zaps we haven’t yet seen.

Predicting the mean zap in unseen data.

Matlab

```
O_zaps = [8 10 10 10 14 14 18 18 18 18 18 22 28 36];

hist(O_zaps, 8:2:36);
set(gca, 'FontSize', 16, 'FontWeight', 'bold');
```



Predicting the mean zap in unseen data.

ZAP

- This is a good start…
- But we know future events will not be *exactly* the same as past events.
- So, the mean zap will not always be: ZAP

- What else might it be?
- ……

Introducing: The Bootstrap!

Courtesy of Rudolph Erich Raspe. Used with permission.

Bootstrapping: Make more samples, measures

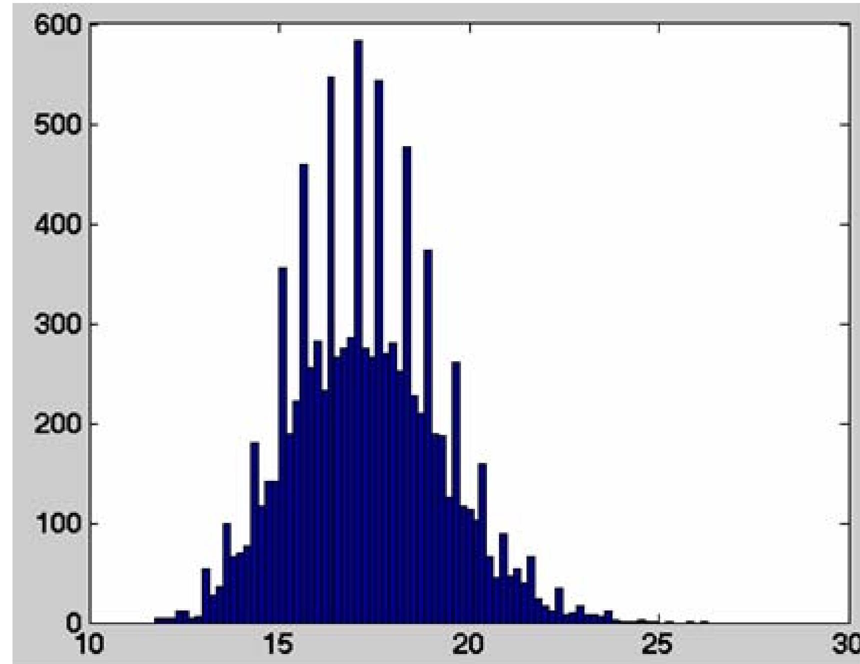
- General reasoning:
  - We will see ‘more of the same’
  - We can produce more of the same to predict the future
  - Compute measure (mean) on more of the same
  - Tabulate the value of the measure.

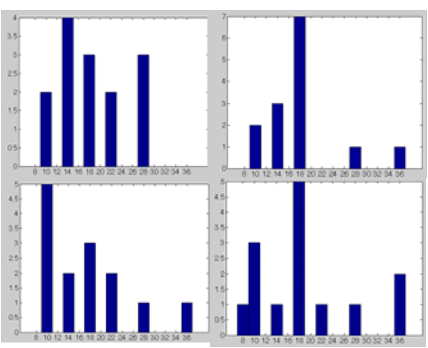
Bootstrapping, more specifically

- We have a sample X containing n observations
- Generate possible future samples:
  - From X draw n times, producing B_1 (another possible sample)
  - Compute measure f() on B1 = M1
  - Repeat # times.

Predicting the mean zap in unseen data.

```
ntimes = 10000;
n = length(O_zaps);

f = @(x)(mean(x));

for i = [1:ntimes]
   B = randsample(O_zaps, n, true);
   M(i) = f(B);
end

hist(M, 80);
```



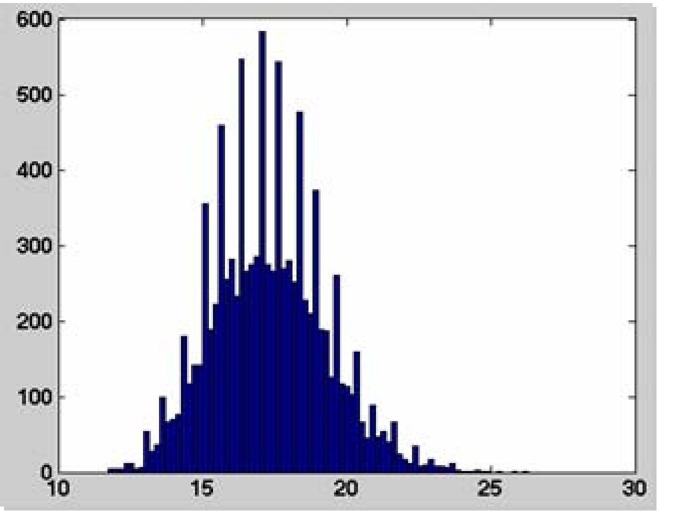


```
function s=randsample(x,n)
      for i = [1:n]
            s(i) = x(ceil(rand()*length(x)));
      end
end
```

- (Don’t use this code – it is really inefficient, consider the Matlab function “bootstrap”)

Predicting the mean zap in unseen data.

- So this represents possible scenarios about what the mean of future data might be.
- Usually we want to say something a bit more concise, like:
  - The mean will be between A and B with confidence P.



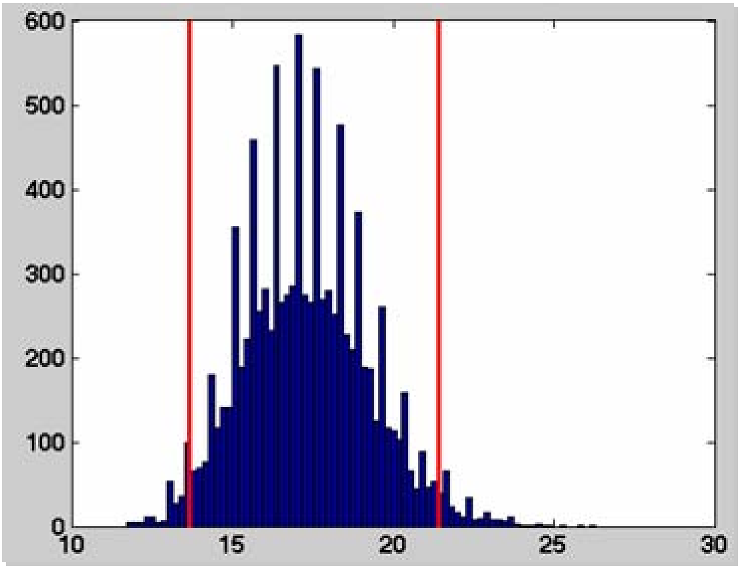
Confidence intervals

- An interval [min to max] which will contain the measure with some level of confidence, P.
  - Confidence as probability
    - Probability as frequency of possible outcomes
- Sort all of our outcomes, consider the bounds of the middle P proportion:

Predicting the mean zap in unseen data.

```
P = 0.95; % confidence level
omitP = 1-P;
lower_bound_percentile = omitP./2;
upper_bound_percentile = 1-omitP./2;
lower_bound_index =
round(lower_bound_percentile*ntimes);
upper_bound_index =
round(upper_bound_percentile*ntimes);

M_sorted = sort(M);

lower_bound = M_sorted(lower_bound_index);
upper_bound = M_sorted(upper_bound_index);

CI = [lower_bound upper_bound]
```

This can all be done with the “quantile” function



With 95% Confidence:
mean zap between 13 and 22

ZAP ZAP

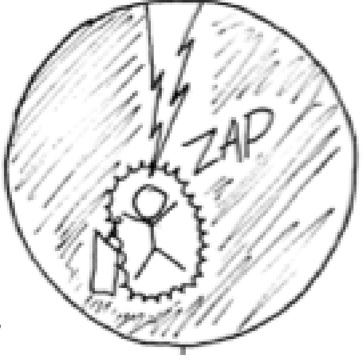
Bootstrapping

- Mean here was a measure.
- You can use *any measure* you like, I won’t judge.

- It’s all good*

- * Some measures are more sensitive to the “Black Swan”

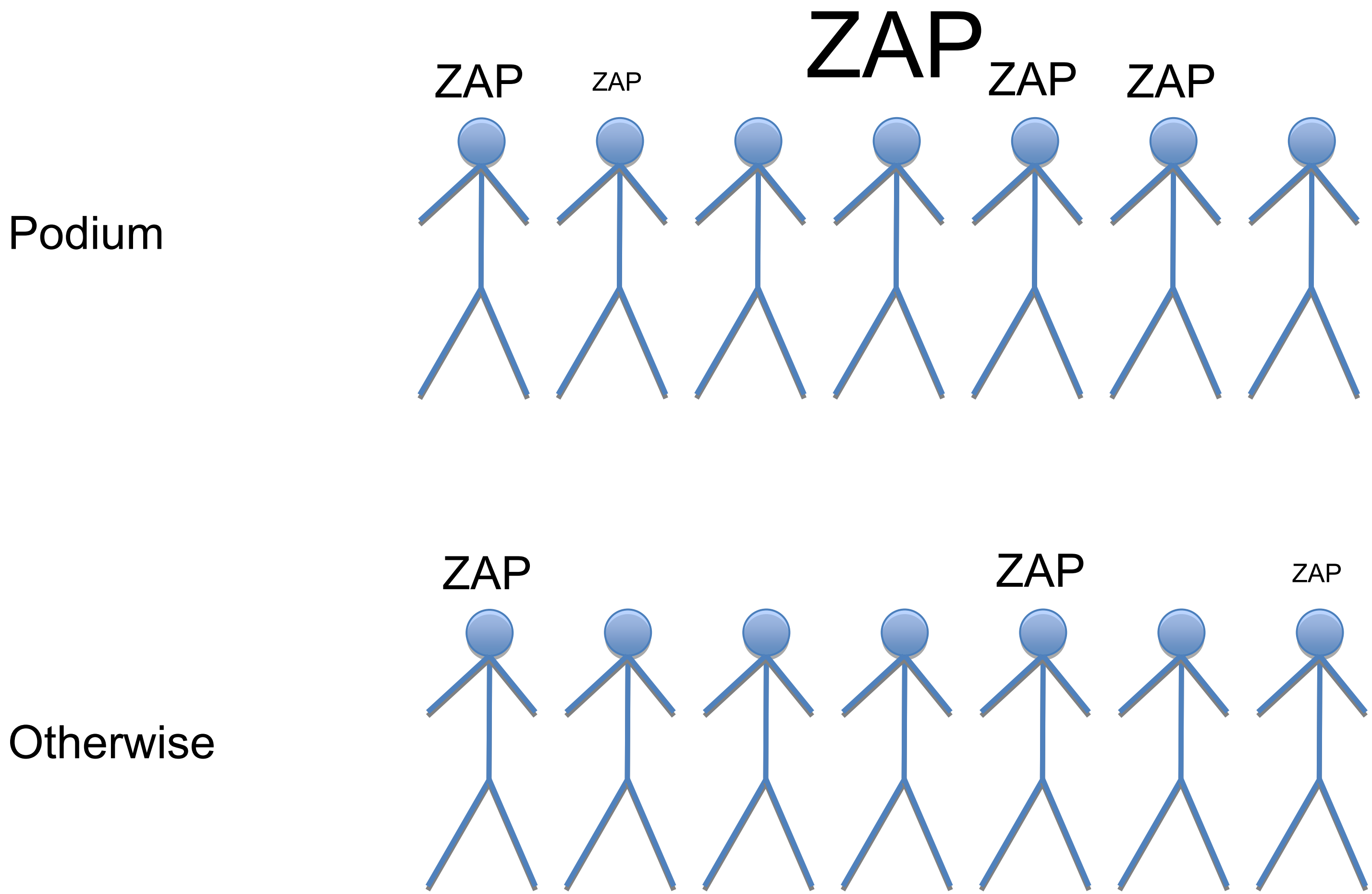
What do we want to know?



Courtesy of xkcd.org

- The mean font size of a zap?
- Do zaps happen more often in this case than otherwise?
- How much bigger are average font sizes at the podium?
- If we got zapped at the podium or somewhere else, which zap would have a bigger font size?

Do zaps happen more often at the podium?



Podium zaps more often than otherwise?

	Zap	No Zap	
Podium	**15**	**5**	75%
Otherwise	8	14	36%

- Well… yes… in this set of observations.
- But we might have observed this difference by chance even if they were the same…

Null Hypothesis Significance Testing

- H_0(null): The effect is 0
  - These groups have the same mean
  - …same frequency of X
  - No correlation is present
- H_1: H_0 is not true.

- Basically: Are these observations *so* improbable under the null hypothesis that we must begrudgingly reject it?

Podium zaps more often than otherwise?

	Zap	No Zap	
Podium	**15**	**5**	75%
Otherwise	8	14	36%

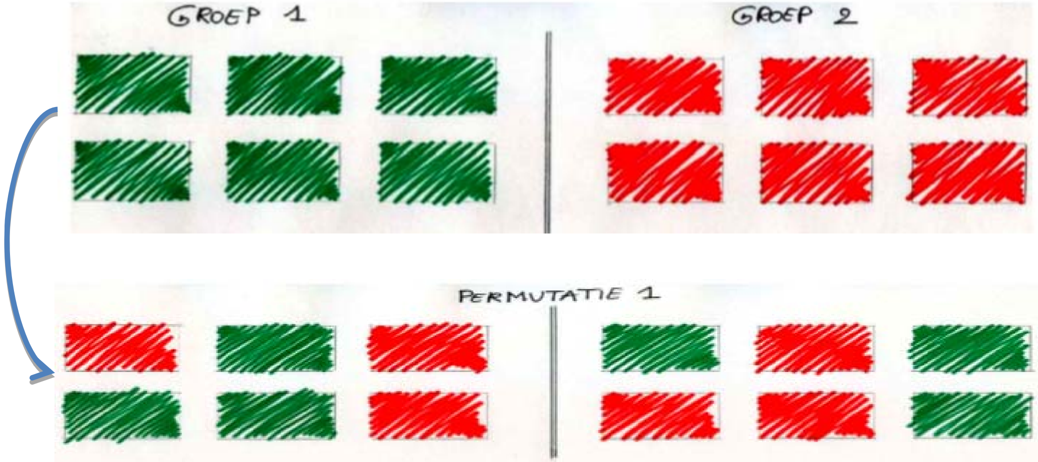
- Well… yes… in this set of observations.
- But we might have observed this difference by chance same…
- How often would a difference at least this big have occurred if these were truly the same? (probability of observing this effect under null hypothesis)

Introducing: Randomization (permutation)

- For most hypothesis tests, null hypothesis is: These things came from the same process.
- So… treat them as such.
- Resample many times from this new combined sample
- Measure the difference of interest in these samples
- See if the difference observed is particularly unlikely

Permutation (simple)

- We have two groups A and B.
- A has n observations, B has m observations.
- Assume they are ‘the same’ (IID), so permute assignments into A and B
  (while maintaining n and m)
- Calculate measure of interest on permutation
- Rinse, repeat.



Podium zaps more often than otherwise?

```
podium = [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 ];
other = [1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0];

f_comp = @(a,b)((sum(a==1)./length(a)) - (sum(b==1)./length(b)));

d_p = f_comp(podium, other);

allobs = [podium, other];
nperm = 10000;

for i = [1:nperm]
   permall = allobs(randperm(length(allobs)));
   perm_podium = permall(1:length(podium));
   perm_other = permall(length(podium)+1:end);

   P(i) = f_comp(perm_podium, perm_other);
end

p = sum(P >= d_p)./length(P);
```

Probability that a difference at least this big would have been observed if these were really ‘the same’? 0.0139

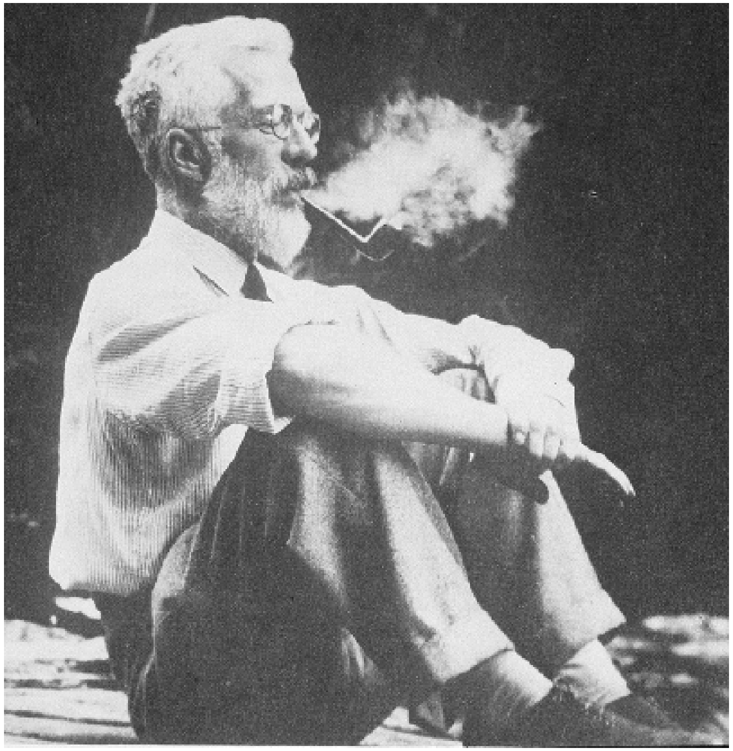
Permutation

- Proportion was a measure here.
- You can use *any measure* you like, I won’t judge.

- It’s all good*.

- * Some measures are more sensitive to the “Black Swan”

How unlikely is *too* unlikely?



... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."…

Fisher, 1926

TABLE III

TABLE OF χ^2

n.	P = ·99.	·98.	·95.	·90.	·80.	·70.	·50.	·30	·20.	·10.	·05.	·02.	·01.
1	·000157	·000628	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·341
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal deviate with unit standard error.

Podium zaps more often than otherwise?

Probability that a difference at least this big would have been observed if these were really ‘the same’? **0.0139**

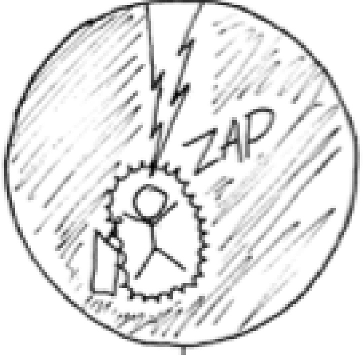
Yes.

“The difference is significant at $p<0.05$.”

“Significant at $p=x$”
This is a little bit weird.

(Talk about tails)

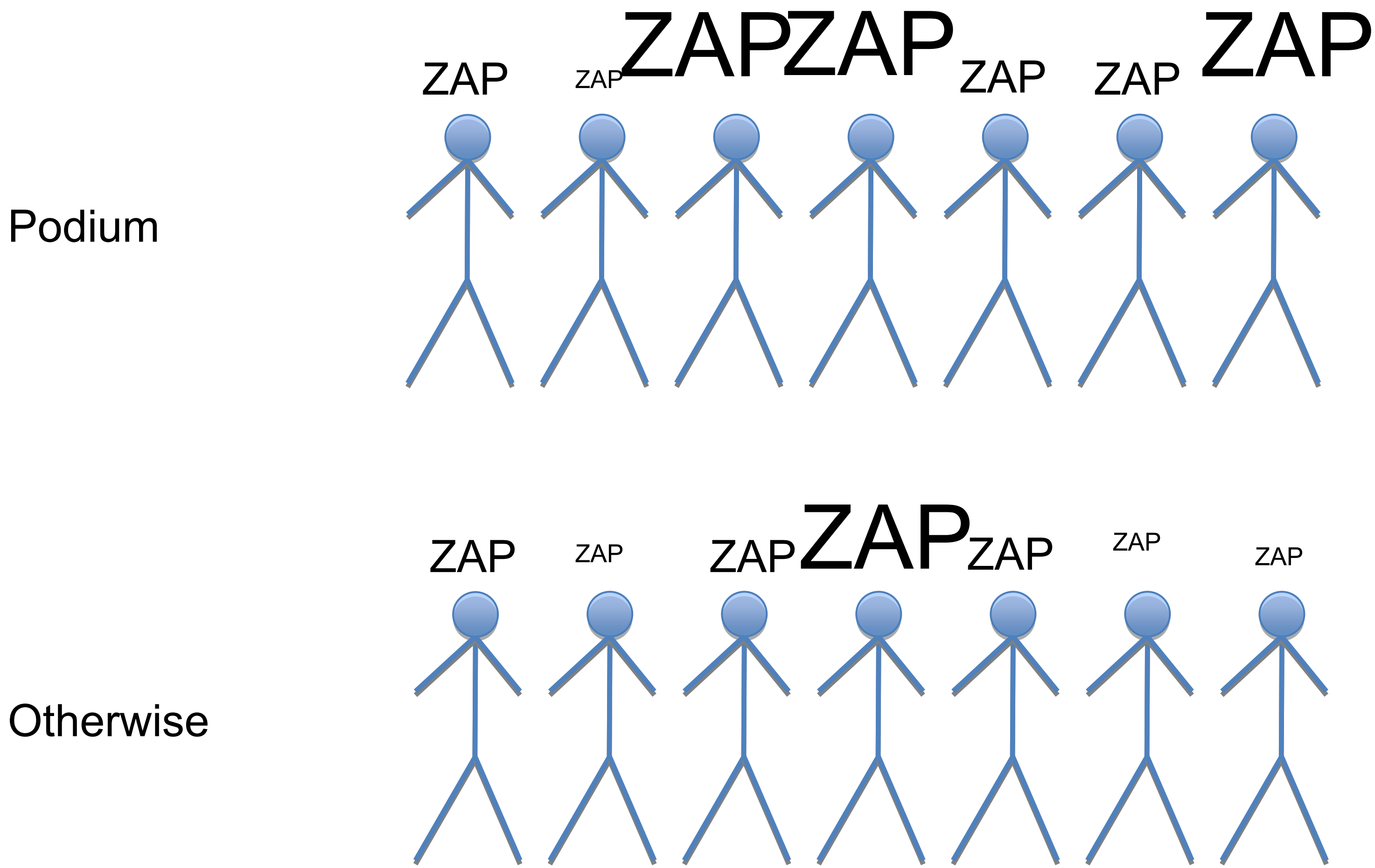
What do we want to know?



Courtesy of xkcd.org

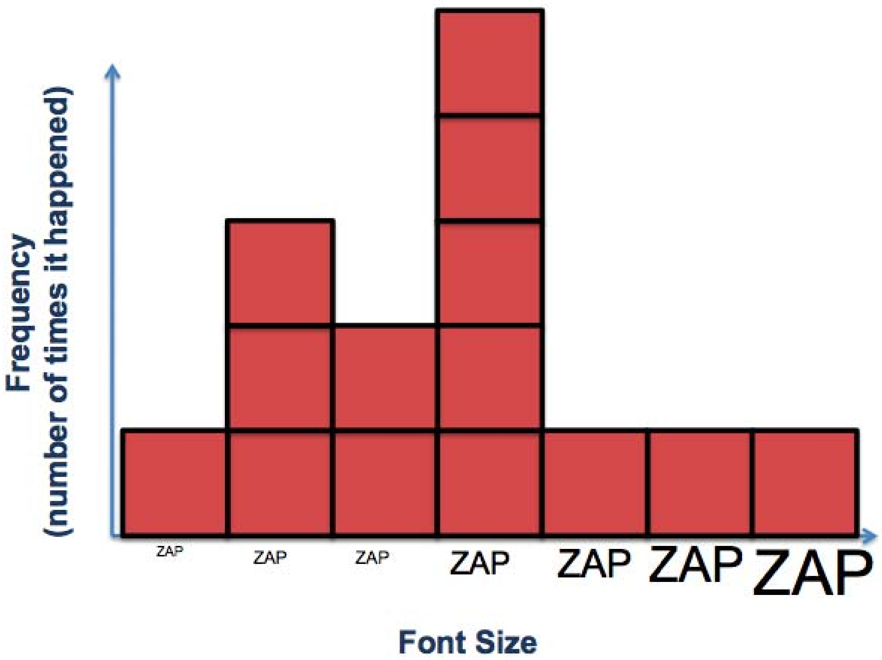
- The mean font size of a zap?
- Do zaps happen more often in this case than otherwise?
- How much bigger are average font sizes at the podium?
- If we got zapped at the podium or somewhere else, which zap would have a bigger font size?

How much bigger are font sizes at podium?

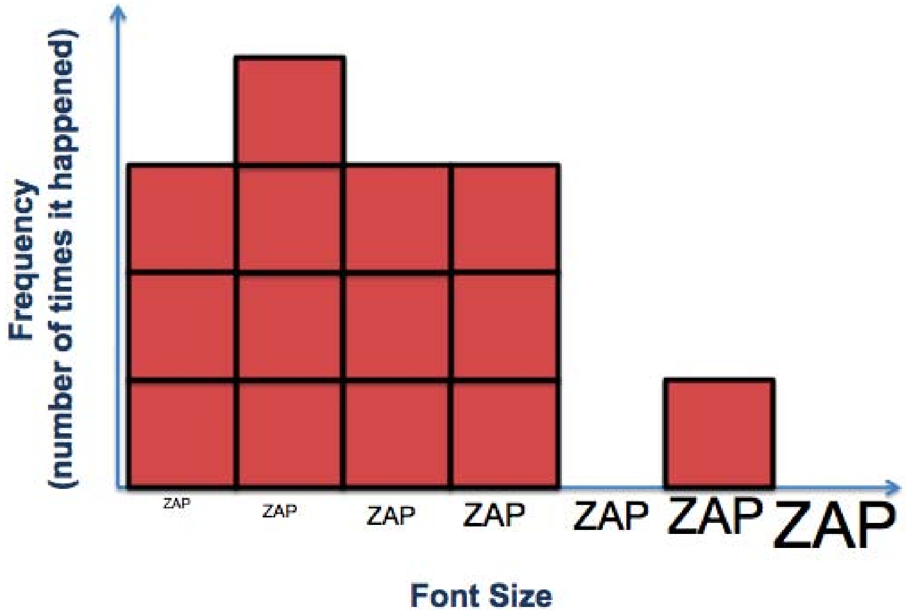


Font sizes observed

Podium



Other



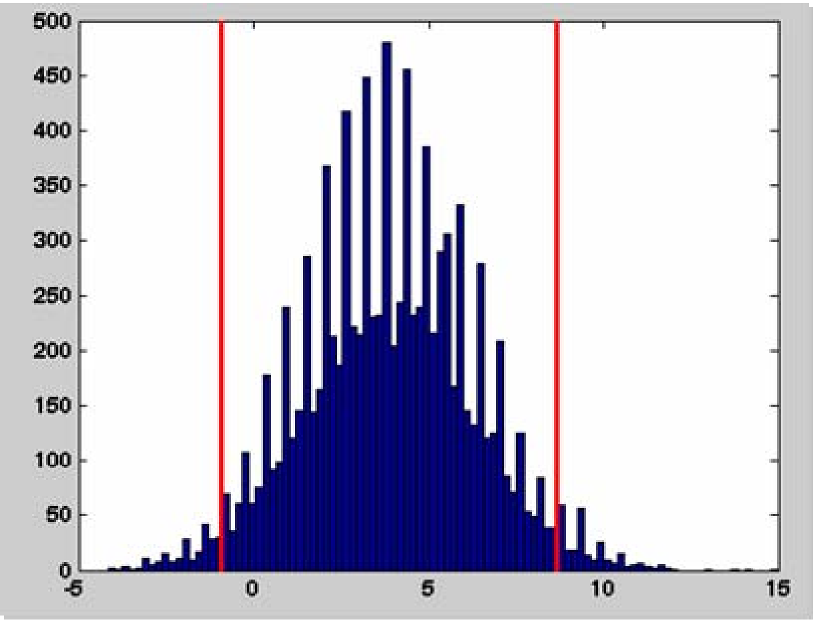
Bootstrapping functions of two samples

- Same thing as bootstrapping one sample.
- Resample each sample
- Compute function of two samples
- Proceed.

Bootstrapping difference of two samples.

```
P_zap = [8 10 10 10 14 14 18 18 18 18 18 22 28 36];
O_zap = [8 8 8 10 10 10 10 14 14 14 18 18 18 28];

f = @(a,b)(mean(a)-mean(b));

nsamp = 10000;
for i = [1:nsamp]
   BP = randsample(P_zap, length(P_zap), true);
   BO = randsample(O_zap, length(O_zap), true);

   M(i) = f(BP, BO);
end

CI = quantile(M, [0.025, 0.975]);
```

Note:
Confidence interval contains zero
This is another way of testing null hypotheses.
(Arguably a much more useful way)



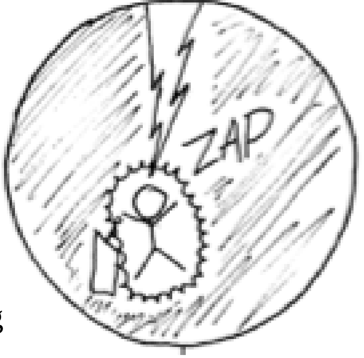
Bootstrapping two-sample measures

- Mean here was a measure.
- You can use *any measure* you like, I won’t judge.

- It’s all good*.

- * Some measures are more sensitive to the “Black Swan”

What do we want to know?



Courtesy of xkcd.org

- The mean font size of a zap?
- Do zaps happen more often in this case than otherwise?
- How much bigger are average font sizes at the podium?
- If we got zapped at the podium or somewhere else, which zap would have a bigger font size?
- Are font sizes more variable at the podium?

Which zap is more likely to be bigger?

- So far we have asked what we might expect of reasonably large samples. If our samples were bigger, we could probably ‘detect’ even smaller changes.
- We don’t care about being able to detect small differences. We often want to know, how much of a difference will it make. Period.
- This is a measure of **effect size**

Dominance (a simple measure of effect size)

- What is the probability that an observation of A will be bigger than an observation of B?
- Choose an A, a B
- Compare
- Repeat

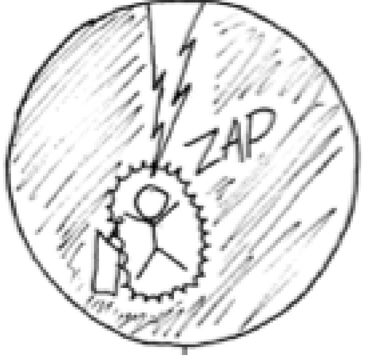
Which zap is more likely to be bigger?

```
f = @(a, b) (a-b);

nsamp = 10000;
for i = [1:nsamp]
   BP = randsample(P_zap, 1, true);
   BO = randsample(O_zap, 1, true);

   M(i) = f(BP, BO);
end

PdO = sum(M>0)./length(M)
OdP = sum(M<0)./length(M)
T = sum(M==0)./length(M)

d = PdO - OdP
```

Podium is bigger	Tie	Other is bigger	dominance
0.58	0.19	0.23	0.35

Podium wins.

What do we want to know?



Courtesy of xkcd.org

- The mean font size of a zap?
- Do zaps happen more often in this case than otherwise?
- How much bigger are average font sizes at the podium?
- If we got zapped at the podium or somewhere else, which zap would have a bigger font size?

What we need

- An assumption of IID observations
- And a computer

What we get

- Predictive distributions of any measure of our choosing:
  - Confidence intervals
  - Significance
  - Effect sizes

What more could we want?

- Ability to deal with "factors"
  - Generally complicated, can do simple cases.
    - Permute within factors
    - (Later) resample residuals (requires more assumptions) (won't get into dealing with multiple factors)
- Work with *really* big datasets.
  - Wrong class, we are doing stuff numerically.

Does the location alter font-size? (one factor)

Podium

Kitchen

Bathroom

Microphone

ZAP
ZAP
ZAP
ZAP

ZAP
ZAP
ZAP
ZAP

ZAP
ZAP
ZAP
ZAP
ZAP

Analysis of within-factor variation

- (I made up this name – there may be an official name out there)
- Define some measure over all three groups, that answers the question:
  "Does this factor alter the observations?"
- Here is an example:
  standard deviation of the mean font-size across different 'levels' of the 'factor'
  (can choose something different, e.g., the range of squared font-sizes across levels)

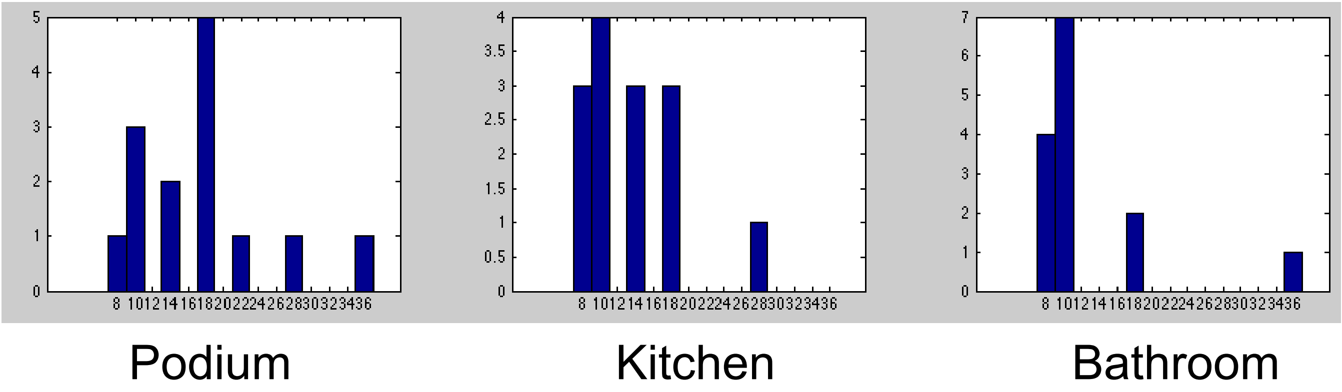
Permute within factors!

- Null hypothesis:
  levels of this factor don’t matter.

- Permute observations across levels
- Build null-hypothesis distribution of this measure.

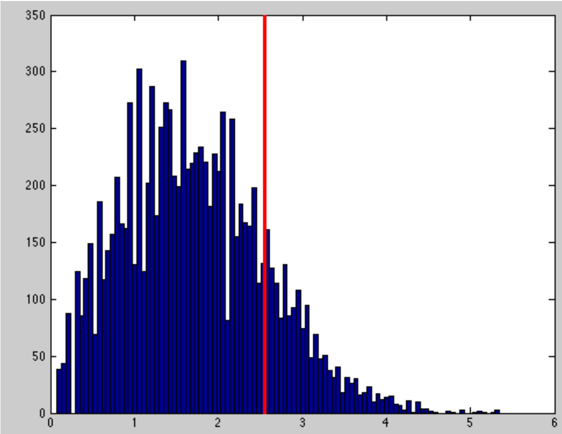
Does the location alter font-size?

```
Z{1,1} = [8 10 10 10 14 14 18 18 18 18 18 22 28 36];
Z{1,2} = [8 8 8 10 10 10 10 14 14 14 18 18 18 28];
Z{1,3} = [8 8 8 8 10 10 10 10 10 10 10 18 18 36];

figure();
for i = [1:3]
   subplot(1,3,i);
   hist(Z{1,i}, [8:2:36]);
end
```



Does the location alter font-size?

```
f_meas = @(a,b,c)(std([mean(a), mean(b), mean(c)]));

Omeas = f_meas(Z{1,1}, Z{1,2}, Z{1,3});

nsamp = 10000;

alldata = [Z{1,1}, Z{1,2}, Z{1,3}];

n1 = length(Z{1,1});
n2 = length(Z{1,2});
n3 = length(Z{1,3});

for i = [1:nsamp]
   P = alldata(randperm(n1+n2+n3));
   P1 = P(1:n1);
   P2 = P((n1+1):(n1+n2));
   P3 = P((n1+n2+1):end);

   M(i) = f_meas(P1,P2,P3);
End

p = sum(M >= Omeas)./length(M)
```



Omeas

P = 0.1654

No
Or “we can’t reject null hypothesis at $p<0.05$”

Permuting within factors

- St.Dev. of Mean across levels was our measure.
- You can use *any measure* you like, I won't judge.

- It's all good*.

- * Some measures are more sensitive to the "Black Swan"

Really Big Limitation

- “Black swan”
  - A general limitation of having incomplete data
- In case of extreme frequentism, even “dirty swans” go ignored.
- We can deal with this (to varying degrees) by specifying beliefs about our ignorance

What more could we want?

- Prettier histograms (more with less)
  - Getting a little Bayesian

- Respect dependencies in data
  - Generally complicated, can do simple cases.

- Make inferences about the world, rather than predicting the outcomes of more samples

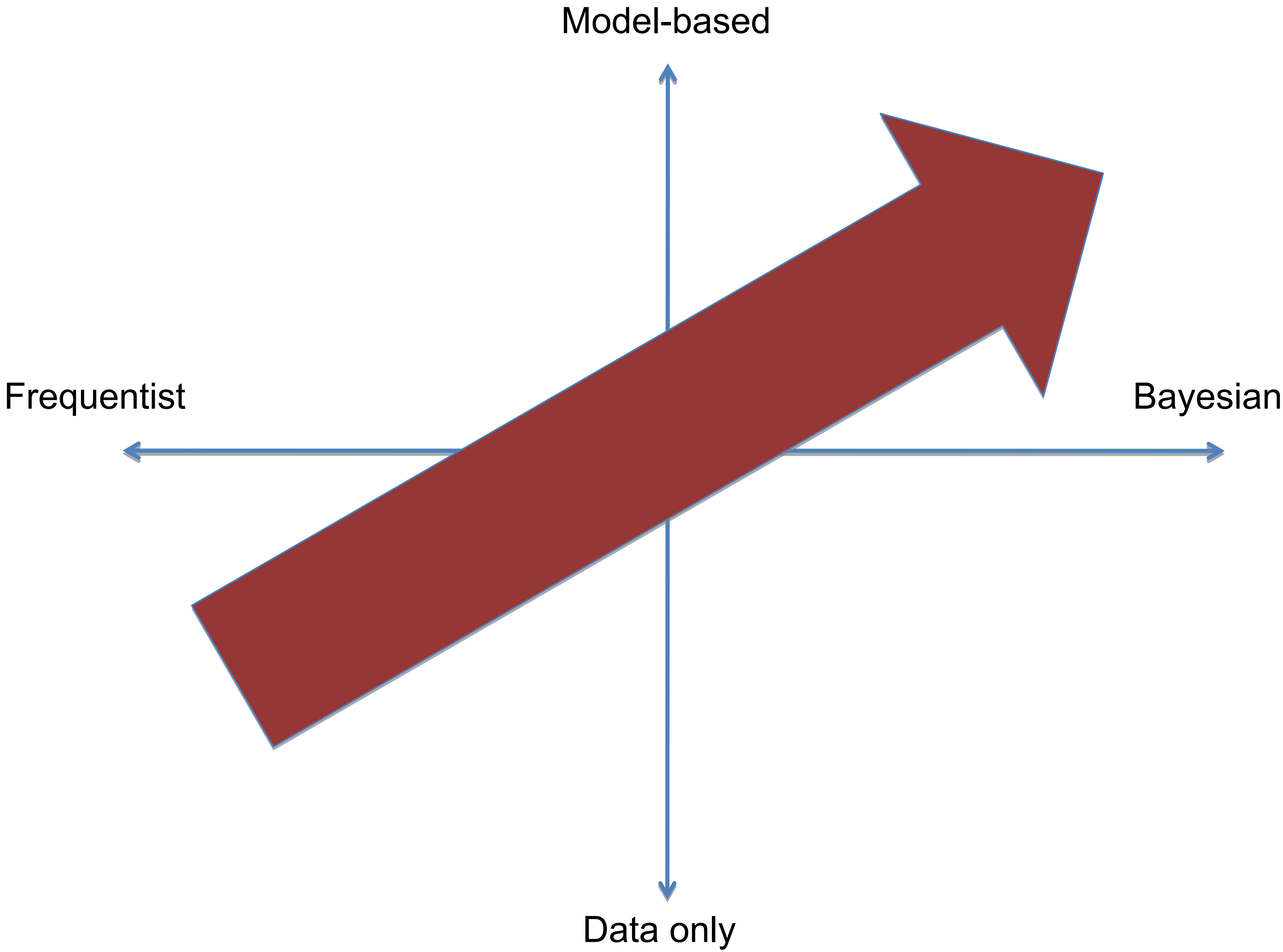
“Yo’ histograms are ugly”

- “I don’t think the real future difference will have those spikes”
- Bayesian!
- New assumption:
  Future data will be
  “more of the same plus noise”
  (kernel density at each data point)

Additional assumptions of ignorance

- Protect against the “black swan” to some extent
- Increase uncertainty
  - Increase range of confidence intervals
  - Decrease the level of significance
- (Note: additional beliefs about underlying distributions [tomorrow] do not just increase uncertainty, and can have worrying effects)

Our class trajectory



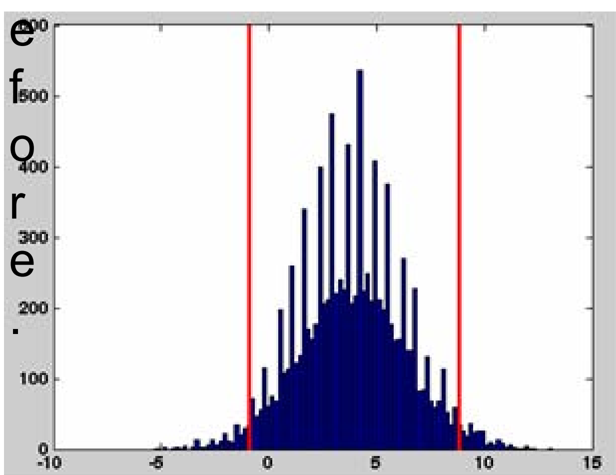
Smoothed bootstrap

- Bootstrap, just as before, but to each draw, add some noise, reflecting our new assumption that future data will be "more of the same plus noise"

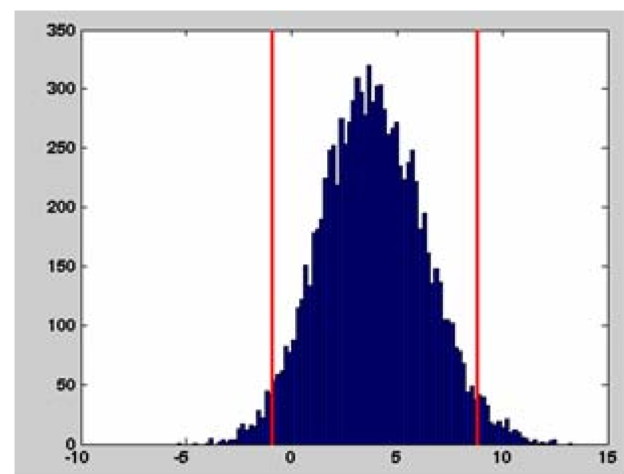
Smoothed Bootstrap

```
for i = [1:nsamp]
   BP = randsample(P_zap, length(P_zap), true)+randn(1, length(P_zap));
   BO = randsample(O_zap, length(O_zap), true)+randn(1, length(P_zap));

   M(i) = f(BP, BO);
end
```

Before.



After.



Hey, this is pretty neat

- I like this Bayesian business.
- What else do I believe about my data that will allow me to **get more from less?**
  - Smoothed bootstrap
  - Resampling residuals
  - Pivoted bootstrap
  - Scaled, pivoted, smoothed bootstrap of residuals…
  - I think there is a distribution in the world…

Residuals are IID; Maybe also Symmetry

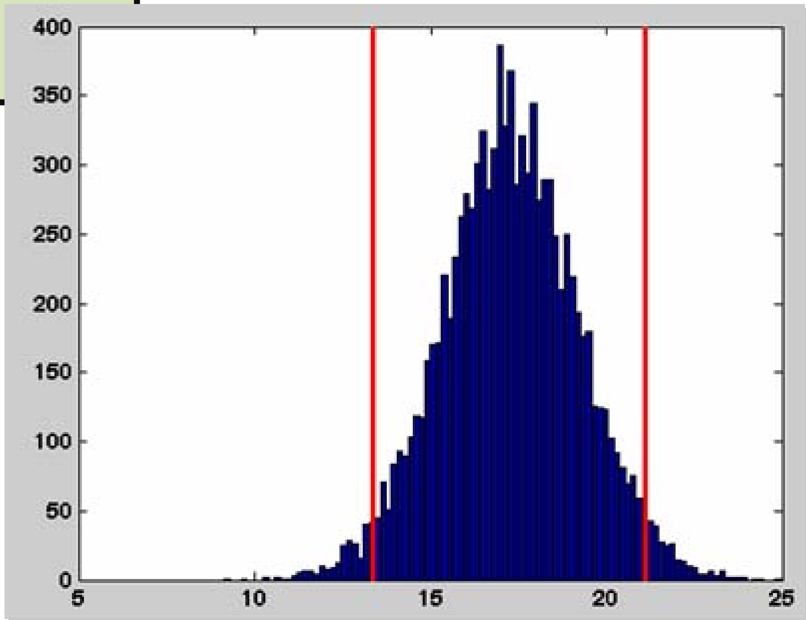
- “More of the same deviations from the mean”
- “More of the same *magnitude* of deviations from the mean”
- Pivoted boostrap

Pivoted bootstrap

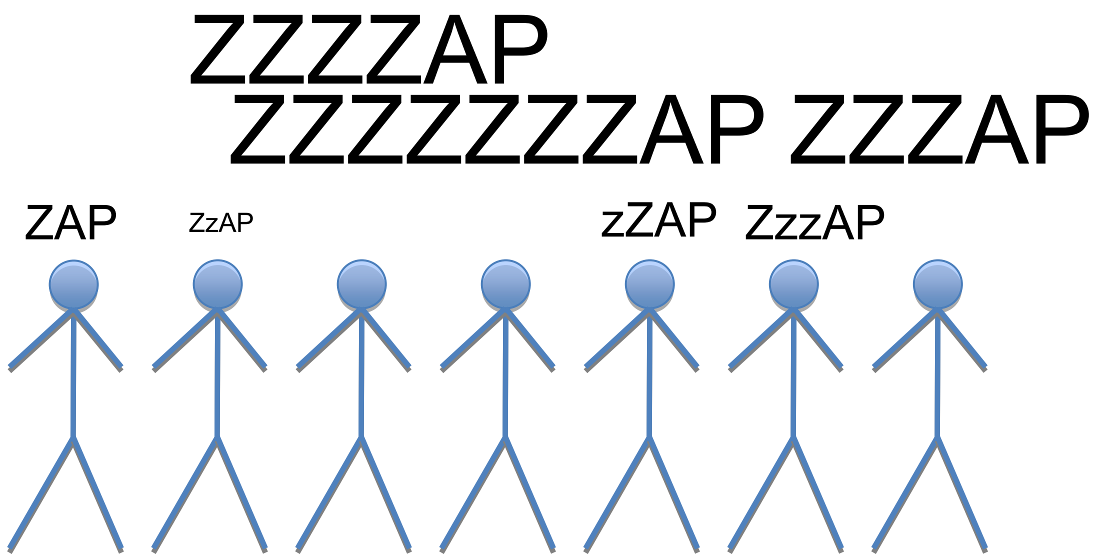
- Compute some measure of central tendency
- Compute deviations from this measure of all observed data
- Bootstrap deviations, and randomly flip sign.
- Add central measure back in to obtain bootstrapped sample
- Compute the bootstrapped measure

Pivoted Bootstrap

```
P_zap = [8 10 10 10 14 14 18 18 18 18 18 22 28 36];

f = @(a,b)(mean(a));
meanP = f(P_zap);

P_zap_dev = P_zap - meanP;

for i = [1:10000];
  B_dev = randsample(P_zap_dev, length(P_zap), true);
  randSign = round(rand(1,length(B_dev))).*2-1;
  B_dev_pivot = B_dev .* randSign;
  B = meanP + B_dev_pivot;

  M(i) = f(B);
end
```

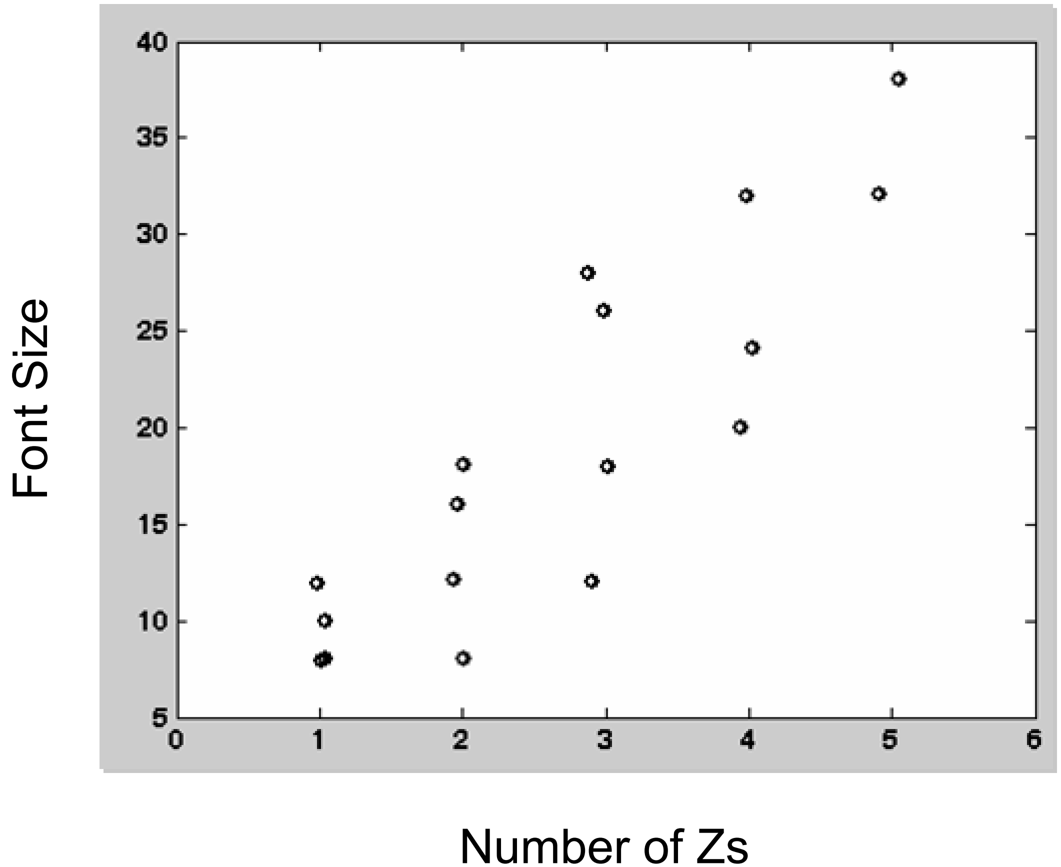


Does number of Zs predict font size?



Does number of Zs predict font size?

```
Ozap = [1 8;
     1 10
     1 8
     1 12
     2 8
     2 12
     2 16
     2 18
     3 12
     3 18
     3 26
     3 28
     4 24
     4 32
     4 20
     5 38
     5 32];
```



Does number of Zs predict font size?

- Measure on the sample of pairs?
- Slope of least-squares regression
  - Why? (Right now, no good reason, but we think it captures something about ‘predicting X from Y’)
  - We could have used some measure on rank orders, etc.

Does number of Zs predict font size?

- Null hypothesis:
  Two dimensions are independent.
- Procedure: resample from them independently to construct new paired sample
- Obtain measure on new sample
- Repeat, build null-hypothesis distribution, etc.

Does number of Zs predict font size?

- Confidence intervals are more useful.
- How do we bootstrap confidence intervals on measures of dependency?
- We often only have one observation at each level of a variable…

- Resample residuals!

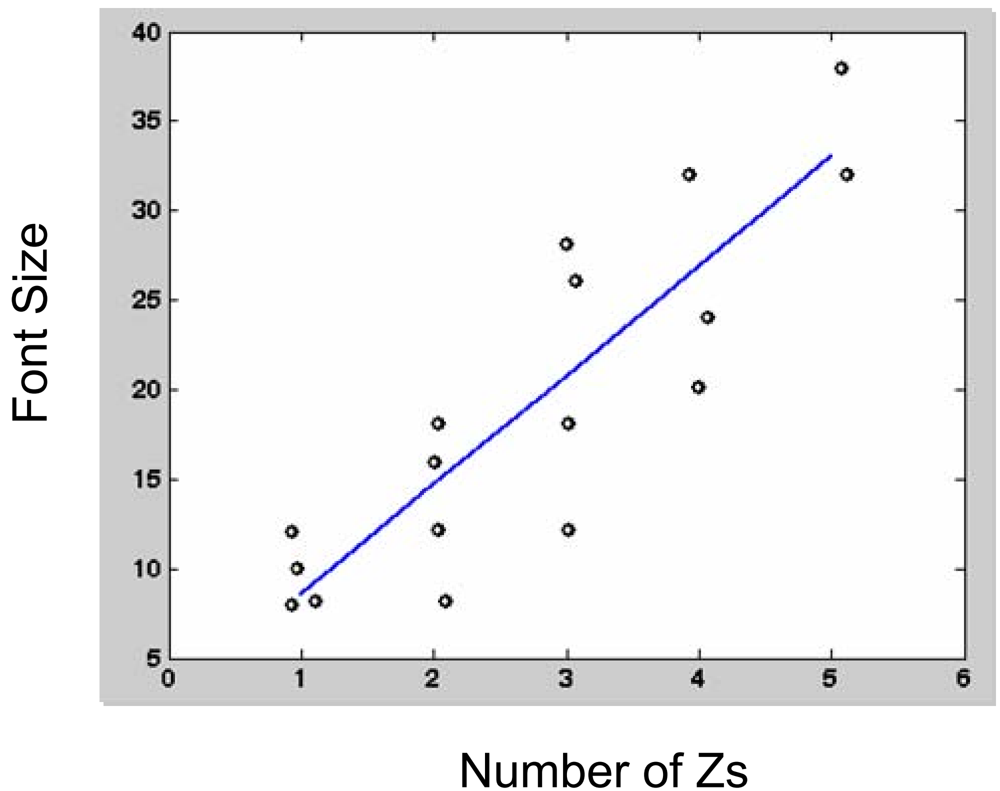
Estimating dependencies in data

- Correlation, regression
- We have a set of paired observations.

- Least squares regression parameters

Does number of Zs predict font size?

```
regression_params = regress(Ozap(:,2), [Ozap(:,1), ones(length(Ozap),1)]);
m = regression_params(1);
b = regression_params(2);

hold on;
plot([1:5], b+m.*[1:5], 'b-', 'LineWidth', 2)
```



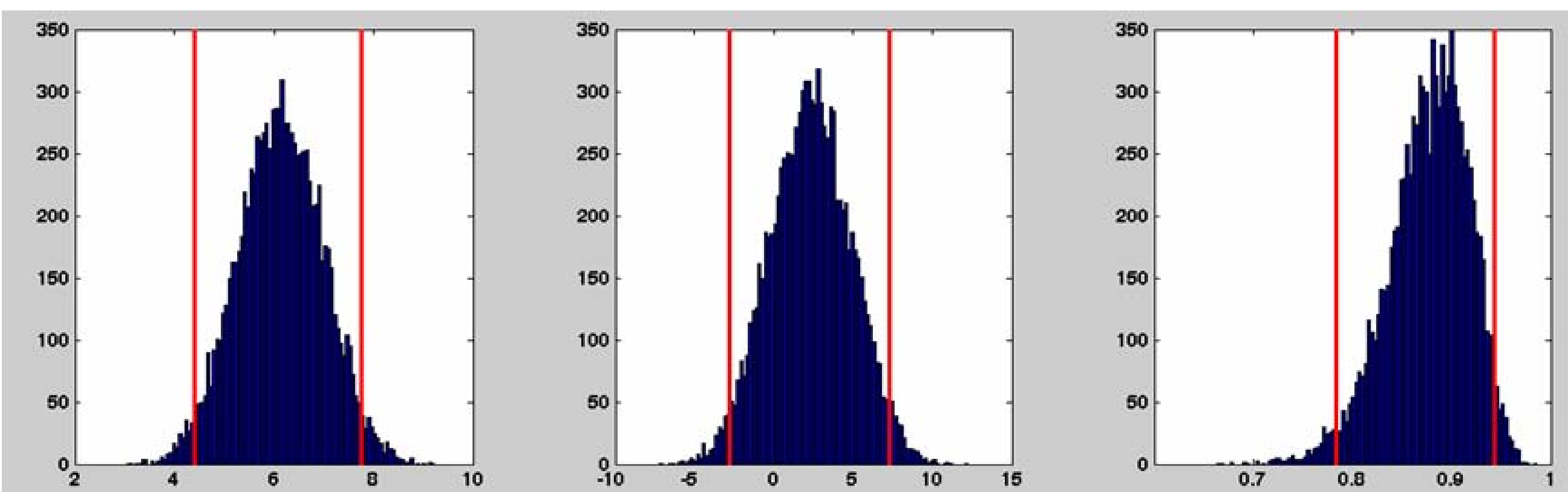
Smoothed, pivoted bootstrap of residuals

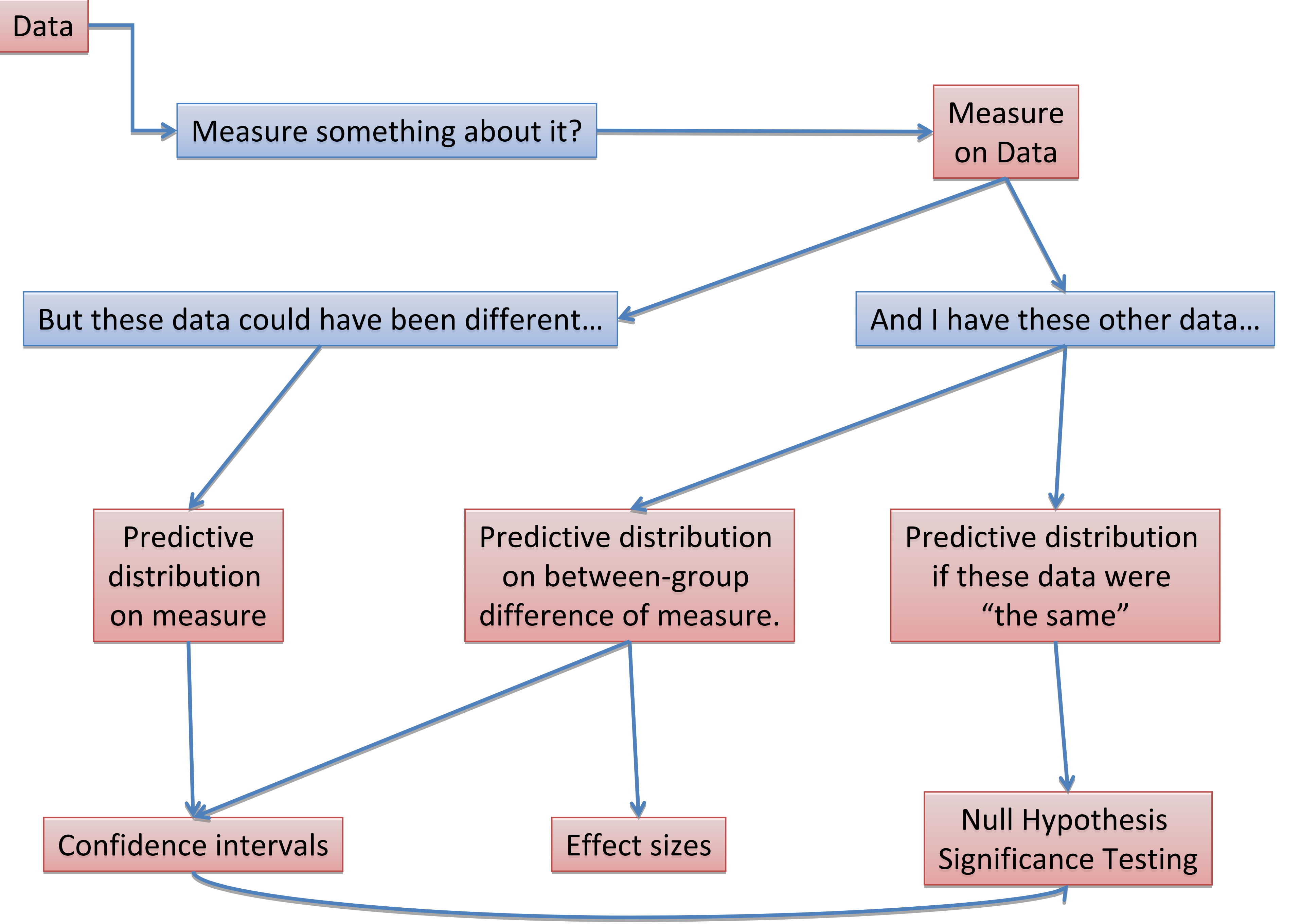
```
res_z = Ozap(:,2) - (b+m.*Ozap(:,1));
for i = [1:10000]
   nz = Ozap(:,1);
   B_res = randsample(res_z, length(nz), true);
   randSign = round(rand(length(B_res),1)).*2-1;
   B_res_piv = B_res .* randSign;
   B_res_piv_smoothed = B_res_piv + randn(length(nz),1);

   B_fs = b + m.*nz + B_res_piv_smoothed;
   regression_params = regress(B_fs, [nz, ones(length(nz),1)]);
   Mm(i) = regression_params(1);
   Mb(i) = regression_params(2);
end
```

Slope

Intercept

Correlation coefficient
(not shown)




Data
Measure something about it?
Measure on Data
But these data could have been different…
And I have these other data…
Predictive distribution on measure
Predictive distribution on between-group difference of measure.
Predictive distribution if these data were "the same"
Confidence intervals
Effect sizes
Null Hypothesis Significance Testing

What we have learned

- Resampling ("more of the same")
- Permutation ("condition assignment is random")
  Null Hypothesis Significance Testing
- Bootstrapping ("more of the same" + measure)
  Confidence Intervals
  - Smoothed ("more of the same + noise")
  - Residuals ("more of the same deviations")
  - Pivoted ("more of the same *symmetric* deviations")
- Dominance to measure effect size
- **Watch out for the black swan!**

Our class trajectory