Contextualizing Coefficients

- Compare the magnitude to other variables, especially variables that are familiar to readers (e.g., age, sex, race). For example "Each additional year of education was associated with an increase of X points in performance on the cognitive test. The magnitude of this association was as large as the predicted difference between cognitive scores of a 75 vs 70 year old."

- Convert the outcome, the exposure, or both to z-scores.
  - Usually not good to present everything this way, but it can be helpful to interpret one or two coefficients.

- Dichotomize your exposure, or describe "The difference in predicted outcome for individuals at the 75th percentile of exposure compared to individuals at the 25th percentile of exposure was…."

Choosing reference groups

- Please make sure dichotomous variables are coded as 0/1.
- Good reference groups:
  - There are a lot of people in the group
  - It represents some sort of standard comparison point
  - It allows you to show the contrast you would like to show.
  - For ex: If the coefficient for men is b=.2 (p=.2) and the coefficient for women is b=.5 (p=.01), then if you code *men* as the reference group, with an interaction between female*attachment, you could get a non-significant main effect for attachment and a non-significant interaction. Technically true, but maybe not the story you'd prefer to present.

Uncertainty

Outline

- Confidence intervals
- Intuitions of bootstrapping
- P-values
- Subgroup effects

Uncertainty

- Most of this class has focused on methods to avoid bias
- But precision of effect estimates is also important
- Intuition:
  - An estimator is biased if the expected value (average) of the estimator does not equal the causal effect
    - Small point of confusion: we can talk about the magnitude of bias or about the qualitative state of being a biased or unbiased estimator
    - A biased but *consistent* estimator gets closer to the truth as you increase the sample size. Sometimes that's good enough.
  - An estimate is imprecise if the estimate would differ if you repeated the whole study – from sample to analysis – exactly the same but with a new set of observations.
- An unbiased but very imprecise estimate may be no use, possibly worse than a precisely estimated estimate w/ a small bias.

Uncertainty

- An estimate is imprecise if the estimate would differ if you repeated the whole study – from sample to analysis – exactly the same but with a new set of observations.
- How can you know how much the estimate would differ if you repeated the whole study with different observations?
  1. Repeat the study with new observations (a bunch of times)
  2. Apply a formula for the sampling distribution of the estimator (if you are lucky enough to know such a formula)

Standard Errors

- Standard error of the sample estimate of the population mean: $\sigma/\sqrt{n}$ or estimated: $s/\sqrt{n}$

- If I drew another sample from the population, a sample with exactly the same n, my estimate of the mean of the population would not be the same.
- I could do this over and over again, drawing a million samples and each time calculating the mean.
- The standard error of the estimated mean describes the distribution of this estimate across my million samples.

- You can redo it over and over again, or estimate it using $s/\sqrt{n}$

Standard Errors

Many statistics we use frequently are popular because they have known sampling distributions, so you can calculate their standard error without redoing the study over and over again

1. Mean (X): SE= $s/\sqrt{n-1}$, var=$s^2/(n-1)$
2. Probability(X=1)=p: SE=$\sqrt{p*(1-p)/(n-1)}$ or =$\sqrt{p*q/(n-1)}$
3. Linear regression coefficient of Y on X:
   - SE(beta)=$\sqrt{\sigma_\varepsilon^2/(n-1)S_X^2}$
   - (the variance of the residuals divided by (n-1)*the sample variance of the exposure)

- Hint: Don't get lost in the details – remember what's in the numerator and what's in the denominator .

Standard Errors

But you may have an estimator for which the sampling distribution is not known, or at least not known by either you or Stata.

For example:

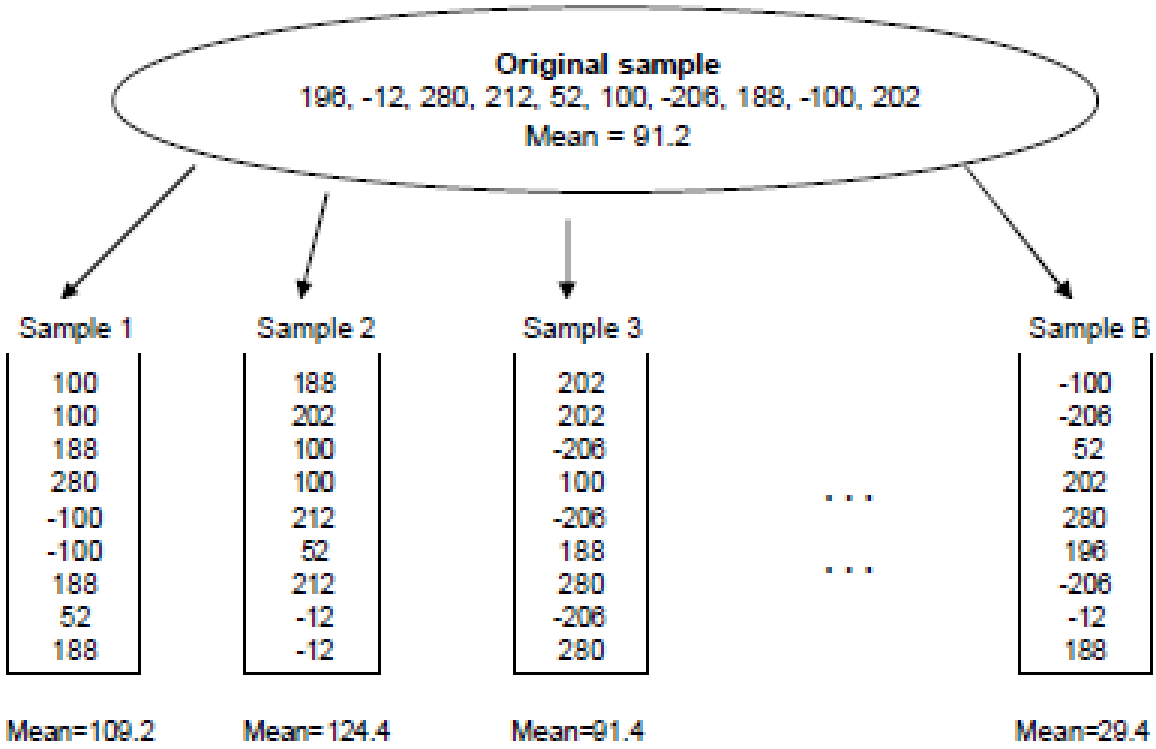
- The 25th percentile minus the 75th percentile
- The difference between two regression coefficients estimated from the same data in non-nested models.
  - Eg immediate risk vs cumulative risk models
- The mean predicted value for the top 20% of the distribution if you estimate among women only versus if you estimate among everyone.

Standard Errors: Bootstrap

- In these cases, the bootstrap is very useful.
- Conceptually, estimate the parameter over and over again in new samples,
- Except instead of actually drawing new samples just a resample of the original sample
- If you resample without replacement, there's no new information!
- If you resample with replacement, some people are represented multiple times, and some people not at all

Bootstrapping

- Useful if you can estimate the parameter, but you don't know how to calculate the standard error for that parameter estimate.
- Resample your original sample with replacement, to create a new population of the same size
- In the new sample, estimate the parameter.
- Repeat this a bunch of times (e.g. 1,000)
- Look at the distribution of your parameter estimate in each of the new samples.
- What is the variance of the estimate across samples?



From Barker, paper PK02

Bootstrapping

- Resample the sample following the same structure used to draw the original sample
- If the original sample was clustered, resample the *clusters*
- In STATA, the command to bootstrap is really straightforward: "bootstrap" or now "bs …. : reg…" or "bs …: blah.do".
- In SAS, you need more code:

From Barker, paper PK02

```
data bootsamp;
  do sampnum = 1 to 1000;
    do i = 1 to nobs;
      x = round(ranuni(0) * nobs);
      set original
          nobs = nobs
          point = x;
      output;
   end;
 end;
 stop;
run;
```

- Then run your analysis with "by sampnum" and output the parameter estimate.

Bootstrapping: issues

- The structure of the bootstrap resampling must match the structure of the *original* sample design. If the original sample design was clustered, the bootstrap must draw a clustered sample.
  - Take a random sample of people, not of the observations on each person
  - Take a random sample of neighborhoods, not of the observations within neighborhoods
  - This means the resamples will not have the same number of observations as the original sample, but will have the same number of clustering unit

Bootstrapping: issues

- Sometimes the bootstrap doesn't work. I don't understand the rules, but ask a stats person before applying it for any specific estimator.
- There is a large technical literature on "bias correction" in the bootstrap. In my experience, it doesn't make a big difference.
    - Take the 2.5th percentile and 97.5th percentile observations and use those as 95% CIs.
    - Take the standard deviation of the effect estimates and use those to calculate the CI
        - If you use this strategy, you have to decide if the center is the median of the bootstrapped estimates or the effect estimate in the primary sample
- If you are calculating a familiar estimator (regression coefficients) with a straightforward standard error, you don't need the bootstrap, but try it a few times to convince yourself that bootstrapping works.

Note the parallel with multiple imputation

1. Estimate the missing value.
2. Estimate the parameter of interest (e.g your regression coefficient)
3. Add uncertainty to your parameter estimate to account for the uncertainty in the imputation of the missing value.

Handling Missing Data: multiple imputation

1. Estimate the missing value.
   - Use the observed values of that variable and all others to estimate the missing value.
   - Estimate it as the predicted value conditional on the other covariates + a random error term, where the random term follows the overall distribution of the variable
   - Repeat this several (e.g., 20) times: each time the "fixed" part is the same, but the random part differs.
   - You now have several (e.g., 20) *possible* data sets

Imputation: Multiple Copies of Dataset

Y	X1	X2	X3
44.61	11.37	178	1
54.3	8.65	156	0
49.87	9.22	.	.
.	11.95	176	1
39.44	13.08	174	1
50.54	.	.	1
44.75	11.12	176	0
51.86	10.33	166	0
40.84	10.95	168	.
46.77	10.25	.	.

_I_	Y	X1	X2	X3
1	44.61	11.37	178	1
1	54.3	8.65	156	0
1	49.87	9.22	181.2	0.23
1	39.97	11.95	176	1
1	39.44	13.08	174	1
1	50.54	9.117	168.2	1
1	44.75	11.12	176	0
1	51.86	10.33	166	0
1	40.84	10.95	168	0.756
1	46.77	10.25	185.9	0.632

_I_	Y	X1	X2	X3
2	44.609	11.37	178	1
2	54.297	8.65	156	0
2	49.874	9.22	137.47	0.0666
2	39.849	11.95	176	1
2	39.442	13.08	174	1
2	50.541	9.9192	162.67	1
2	44.754	11.12	176	0
2	51.855	10.33	166	0
2	40.836	10.95	168	0.2288
2	46.774	10.25	184.83	0.0998

From: CFDR, BGSU

Imputation: Iterative

Y	X1	X2	X3
44.61	11.37	178	1
54.3	8.65	156	0
49.87	9.22	.	.
.	11.95	176	1
39.44	13.08	174	1
50.54	.	.	1
44.75	11.12	176	0
51.86	10.33	166	0
40.84	10.95	168	.
46.77	10.25	.	.

_I_	Y	X1	X2	X3
1	44.61	11.37	178	1
1	54.3	8.65	156	0
1	49.87	9.22	181.2	0.23
1	39.97	11.95	176	1
1	39.44	13.08	174	1
1	50.54	9.117	168.2	1
1	44.75	11.12	176	0
1	51.86	10.33	166	0
1	40.84	10.95	168	0.756
1	46.77	10.25	185.9	0.632

_I_	Y	X1	X2	X3
2	44.609	11.37	178	1
2	54.297	8.65	156	0
2	49.874	9.22	137.47	0.0666
2	39.849	11.95	176	1
2	39.442	13.08	174	1
2	50.541	9.9192	162.67	1
2	44.754	11.12	176	0
2	51.855	10.33	166	0
2	40.836	10.95	168	0.2288
2	46.774	10.25	184.83	0.0998

From: CFDR, BGSU

- Each imputation requires an assumed covariance matrix or regression model for the “fixed” part of the prediction, based on all the observed variables.
- Once you have imputed, your best estimates of these coefficients will change.
- So the computer iterates: estimate the covariance matrix, fill in the data, re-estimate the covariance matrix, fill in new data, re-estimate the covariance matrix… until it’s stable.

Handling Missing Data: multiple imputation

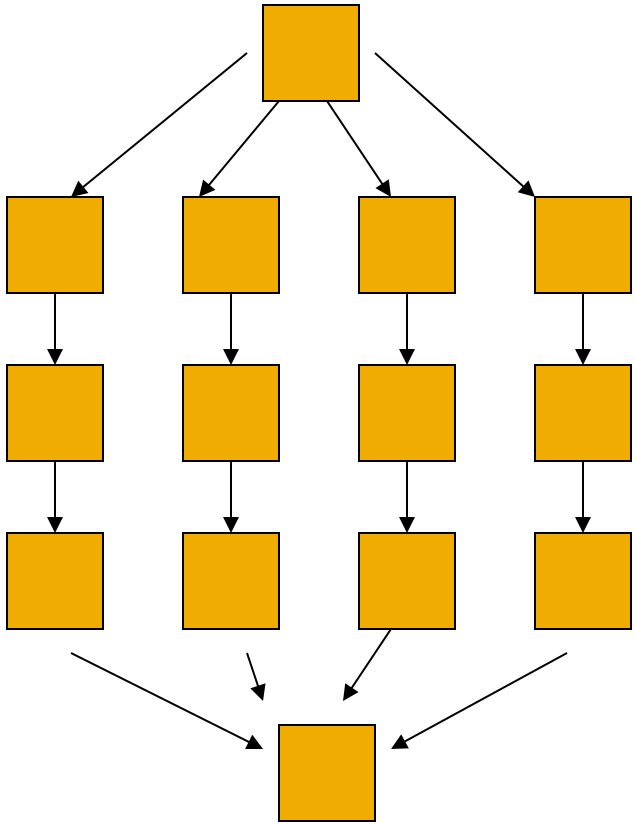
1. Estimate the missing value.
   - You now have several (e.g., 20) *possible* data sets
2. Estimate the parameter of interest (e.g your regression coefficient)
   - Derive one parameter estimate in each of the possible data sets.
   - Now you have a distribution of estimates, some bigger, some smaller
   - Within each data set, there are confidence bounds around the parameter estimate
3. Add uncertainty to your parameter estimate to account for the uncertainty in the imputation of the missing value.

Handling Missing Data: multiple imputation

1. Estimate the missing value.
   - You now have several (e.g., 20) *possible* data sets
2. Estimate the parameter of interest (e.g., your regression coefficient)
3. Add uncertainty to your parameter estimate to account for the uncertainty in the imputation of the missing value.
   - Combine the two sources of uncertainty: within data set variance in your estimate + between data set variance in your estimate

Multiple Imputation: a really nice picture

Rubin 1987



From: Von Hippel

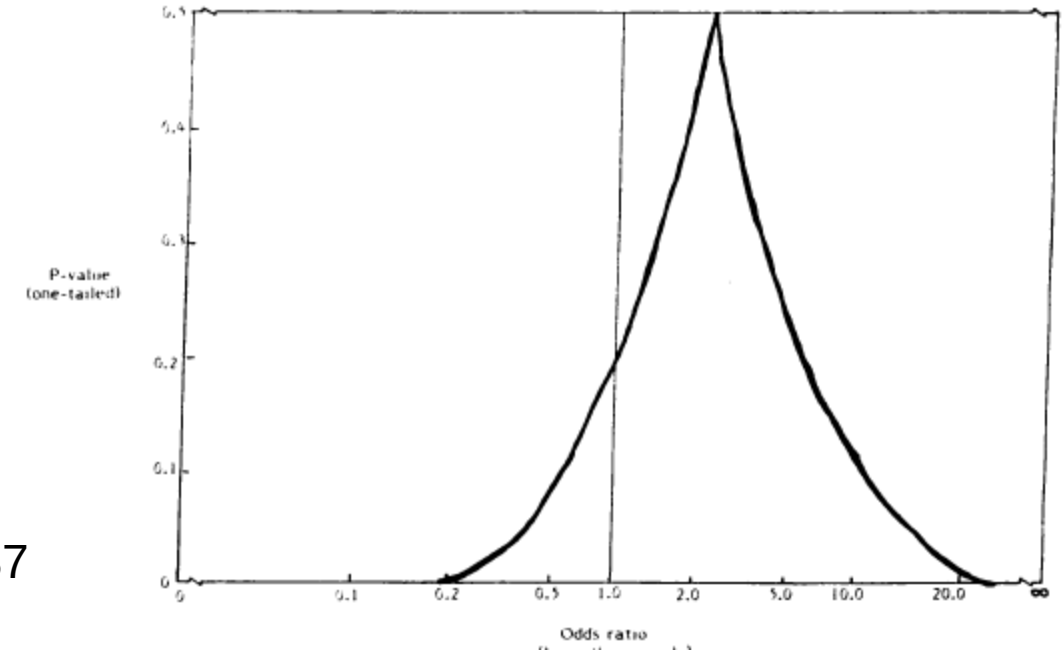
Handling Missing Data: multiple imputation

- This is automated in SAS and Stata
- Some people hate it because it is "making up data"
- Sometimes reviewers want to see it.
- If there is only a small fraction of missing data, it probably won't matter.
- If there's a lot of missing data, or **highly scatter-shot** missing data, it can be very helpful

P-values:

- The p-value > .05, therefore, $\beta=0$
- The 95% CI may include the null, but does it also include values of substantive/clinical importance?
- Large p-values *might* reflect a true effect that is null or close to null.
- But p-values are also influenced by the sample size, and the other sources of variance in the outcome.

- Look at the whole CI, and recognize that your data would be consistent with values across the range.
- Do not simply use the CI as a more irritating way to test p>.05



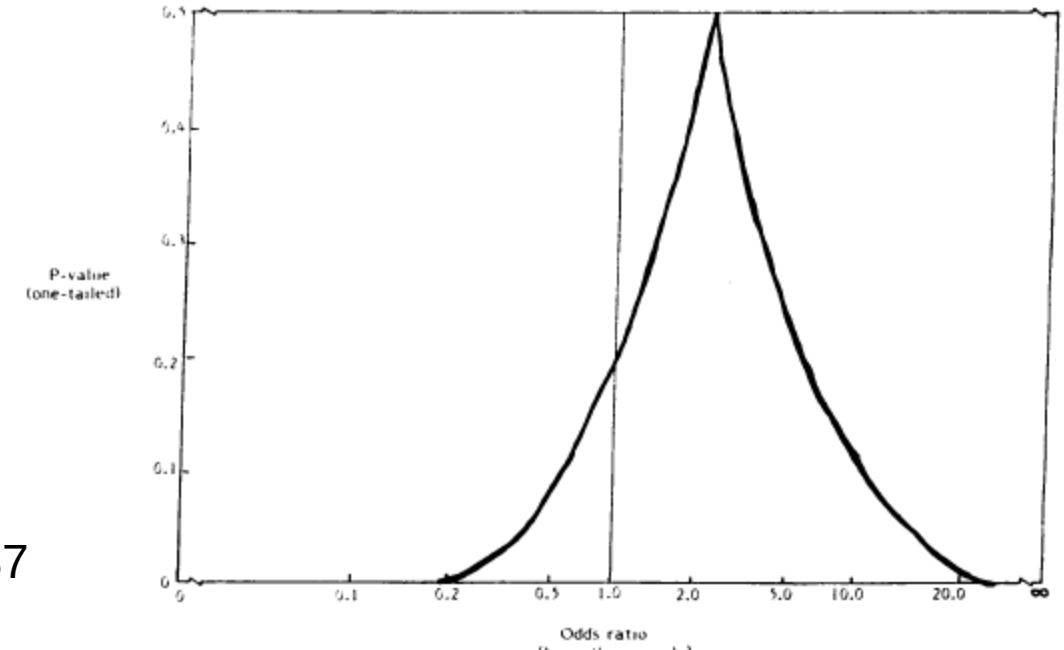
From: Poole, Beyond CIs, 1987

P-values: some common traps to avoid

- The p-value > .05, therefore, $\beta=0$
- The 95% CI may include the null, but does it also include values of substantive/clinical importance?
- Large p-values *might* reflect a true effect that is null or close to null.
- But p-values are also influenced by the sample size, and the other sources of variance in the outcome.

- Look at the whole CI, and recognize that your data would be consistent with values across the range.
- Do not simply use the CI as a more irritating way to test p>.05

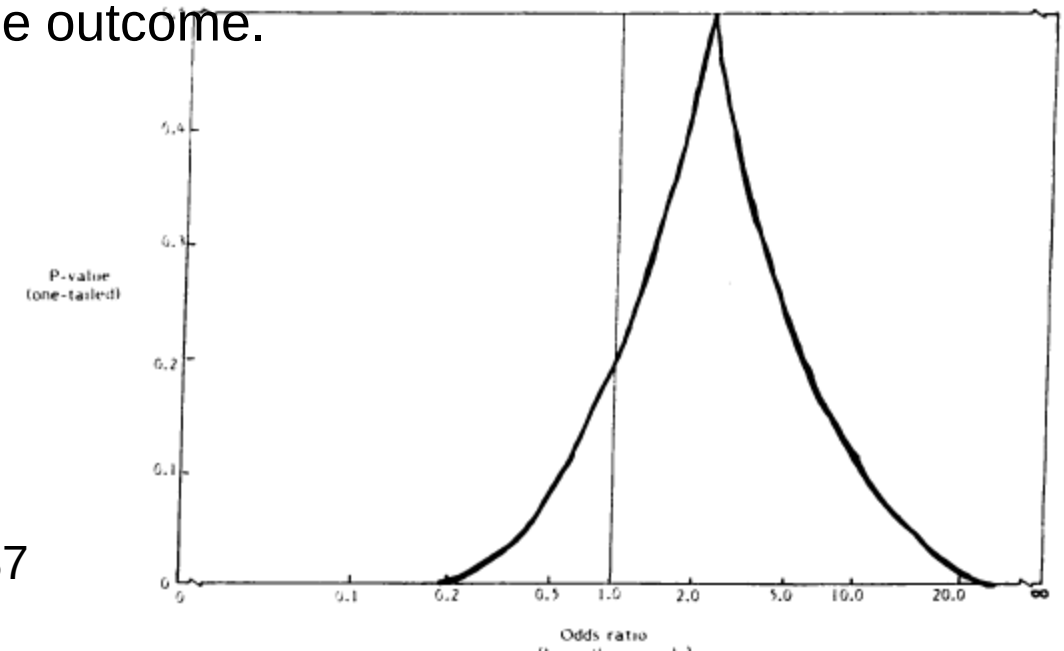
From: Poole, Beyond CIs, 1987



P-values: some common traps to avoid

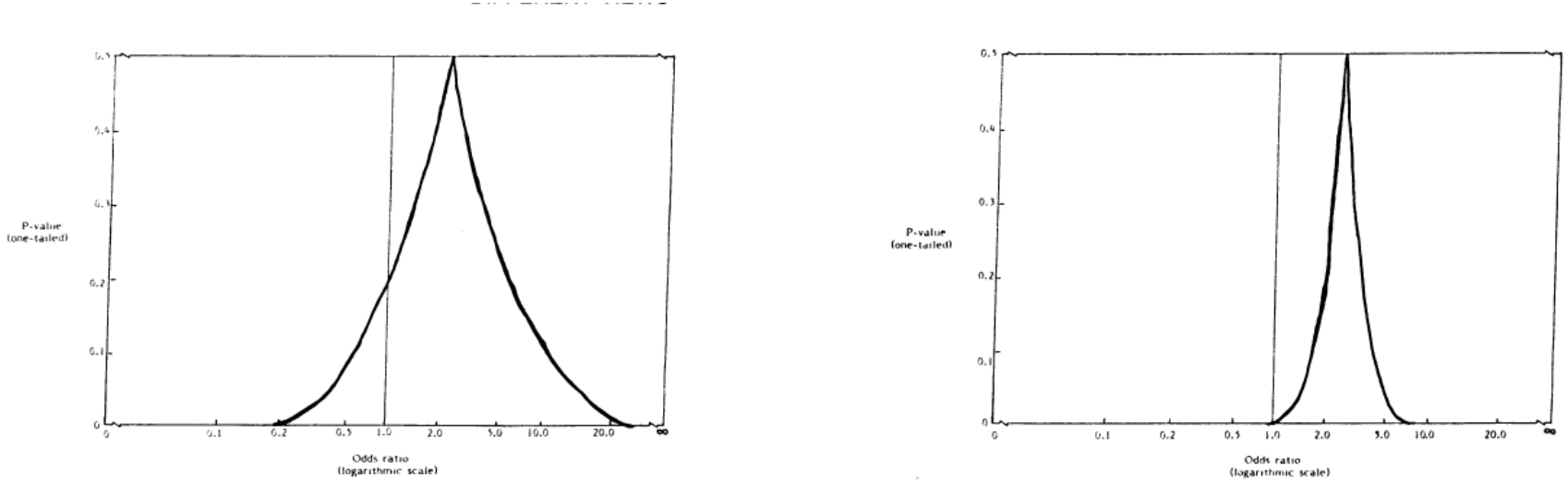
- ~~The p-value > .05, therefore, β=0~~ NO NO NO This is so insidious and tempting…
- The 95% CI may include the null, but does it also include values of substantive/clinical importance?
- Large p-values *might* reflect a true effect that is null or close to null.
- But p-values are also influenced by the sample size, and the other sources of variance in the outcome.
- Look at the whole CI, and recognize that your data would be consistent with values across the range.
- Do not simply use the CI as a more irritating way to test p>.05

From: Poole, Beyond CIs, 1987



P-values: some common traps to avoid

- The p-value > .05, therefore, $\beta=0$



In these two situations, the point estimate is the same, but there is more precision in the right hand panel than the left.

From: Poole, Beyond CIs, 1987

P-values: some common traps to avoid

- Situations when this is tempting:
  - $E(BMI)=b_0+b_1*Mom\_Ed$
  - b_1=-2 95% CI: (-.08, -3.92)

  - $E(BMI)=a_0+a_1*Mom\_Ed+a_2*Own\_Ed$
  - a_1=-1.8 95% CI: (0.20, -3.80)
  - a_2=-4 95% CI: (-3, -5)

Does mother's education have an effect on BMI independent of own education?

P-values: some common traps to avoid

- Situations when this is tempting:
  - E(BMI)=b_0+b_1*Mom_Ed among women
  - b_1=-2 95% CI: (-.08, -3.92)
  - E(BMI)=a_0+a_1*Mom_Ed among men
  - a_1=-1.8 95% CI: (0.20, -3.80)

  Is the effect of mother's education greater among men than among women?

Sample Designs

- Snowball/convenience
- Simple random sample
- Clustered sample
- Stratified sample

- Ideas from sampling extend to inverse probability weighting, with the idea that in sampling you apply weights to recreate the original target population and in IPW you apply weights to recreate the original unselected population or to create a population with no confounding.

- Also, capture recapture…

Conclusions: Causal Inference

- Doll & Hill: only temporal order essential, the rest good for clear thinking
- Cook and campbell
  - Internal validity, external validity, construct validity
  - sources of bias (history, maturation, regression to the mean)
  - Approaches when observational studies match trials
- Counterfactuals and DAGs

Conclusions: data source tradeoffs

- Design tradeoffs (typically)
  - Big sample == worse measures
  - Better measures == smaller, less representative sample
- Surveillance data (vital stats, census)
- Major surveillance studies, NHANES, BRFSS
- Actively enrolled and followed cohorts w/ clinic visits
- Actively enrolled and followed community-based cohorts
- Passive data sources, EHR records
- Clinic based samples with active enrollment
- Case registries

Conclusions: Designs to help identification

- Mixed models
  - Some help with identification, because you can account for compositional differences
  - Mostly help with asking multilevel questions
  - For change over time (aging, growth, development, recovery)
  - For place level clustering
- Quasi-randomization
  - within-person changes
  - IV
  - RDD

Conclusions: Selection problems

- Selection arises from:
  - Eligibility for the study
  - Study enrollment
  - Study dropout
  - Mortality
- Selection processes can compromise external validity (the effect of X on Y differs for people in your study than for people outside of your study)
- Selection processes can compromise internal validity (people in your study with and without X are not exchangeable for Y)
- If you can model the selection process, you can account for it, but you have to understand what influences selection.
- IPW is a common approach to accounting for selection

Conclusions: Uncertainty

- You nearly always see only 1 iteration of your study, 1 sample, 1 set of sampling peculiarities, 1 version of the analyst making his or her innumerable arbitrary decisions.
- Seek replication and honest accounts of uncertainty
  - Uncertainty within studies
  - Uncertainty between studies
  - Both contribute to uncertainty
- Confidence intervals are usually more informative than p-values