Final project

- Read recent issues of a journal in your field (e.g., Am J Epi; Epidemiology; Neurology; Pediatrics). Choose an article that addresses a research topic of interest to you using an observational study design. Summarize the study design, sampling, and analytic plan used in the paper. Draw a DAG corresponding with the author's implicit or explicit causal hypotheses and specify which part of the DAG they are testing in this article. Review their handling of statistical conclusion validity, internal validity and external validity. Describe what you consider the greatest weaknesses of the paper. (This should be ~2000 words).

- Write the methods section for an alternative study that would address the same research topic with either a stronger design or a design that addressed the most important limitations of the initial study. This should be ~1500-2000 words.

Logistic Regression ?

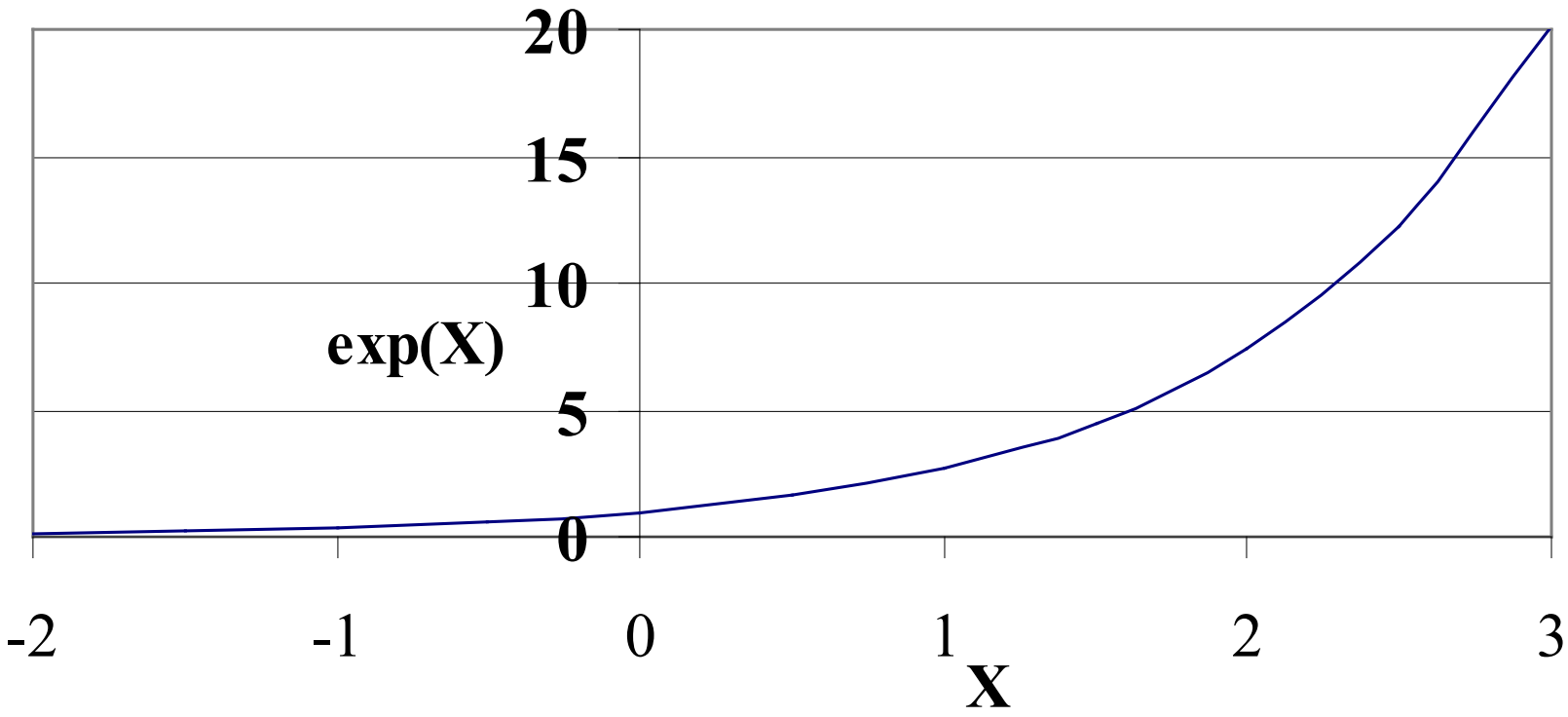
Probabilities vs Odds

- How does the odds relate to the probability?
- Odds=probability/(1-probability)
- Probability=Odds/(1+Odds)
- How can you remember this?
  - "The odds are 50/50" means the same as "the odds are 1"
- P=.01, Odds=.01/.99=.010
- P=.05, Odds =.05/.95=.053
- P=.10, Odds =.10/.90=.11
- P=.25, Odds =.25/.75=.33
- P=.50, Odds =.50/.50=1
- Very similar for rare events, but not for common events.

Probabilities vs Odds

- If you know the probability, you can calculate the odds, and vice versa
- If you know the probability, you can calculate the logit=ln(p/(1-p)
- e~=2.7
- e^2=7.4
- $e^{0.5}$=1.65
- ln(k)=what you exponentiate to get k
- ln(1.65)=0.5
- ln(7.40)=2
- ln(2.70)=1

Plot of exp(X)



Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Probability ranges from 0 to 1

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

$E(Cognitive\_Impairment)=b_0+b_1*Male+b_2*Age+b_3*Male*Age$

$b_0=.15$ $b_1=-.3$ $b_2=.1$ $b_3=-.05$

What is the predicted probability of cognitive impairment for women age=0?

What is the predicted probability of cognitive impairment for women age=20?

Using probability as the dependent variable in a linear model can lead to out-of-possible-range predictions.

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^l/(1+e^l)$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
0.05	0.05	-2.94	0.05	0.05
0.25	0.33	-1.10	0.33	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Ln(odds) ranges from negative infinity to positive infinity

Probabilities vs Odds

probability	odds	ln(odds)	exp(ln(odds))	$e^{l}/(1+e^{l})$
0.00	0.00	!!!		
0.01	0.01	-4.60	0.01	0.01
[illegible]	[illegible]	[illegible]	[illegible]	0.05
[illegible]	[illegible]	[illegible]	[illegible]	0.25
0.50	1.00	0.00	1.00	0.50
0.75	3.00	1.10	3.00	0.75
0.95	19.00	2.94	19.00	0.95
0.99	99.00	4.60	99.00	0.99
1.00	!!!			

Let's REGRESS!!

Ln odds ranges from – infinity to + infinity

Basic logistic model

- Logit(p)= log(p/(1-p)) =α+βX
- α and β are unknown parameters to be estimated from the data (e.g. with MLE).
- This model assumes that the logit increases linearly with X.
- If that is not true, the equation is mis-specified (or possibly mispecified)

- probability=odds/(1+odds)

Basic logistic model

- **Logit(p)= ln(p/(1-p)) =α+βX**

- p/(1-p)=exp(α+βX)Exp — Exponentiate both sides
- p=(1-p)*exp(α+βX) — Multiply both sides by (1-p)
- p=exp(α+βX)-p*exp(α+βX) — Multiply through the (1-p)
- p+p*exp(α+βX)=exp(α+βX) — Add p*exp(a+bx) to both sides
- p*(1+exp(α+βX))=exp(α+βX) — Factor out p
- p=exp(α+βX)/ (1+exp(α+βX)) — Divide both sides by (1+exp(a+bx))
- **$p=e^{l}/(1+e^{l})$**

Basic logistic model

- Logit(p)= ln(p/(1-p)) =$\alpha+\beta X$
- β is the ln of the odds ratio
- Exponentiate β to get the odds ratio
- If X is dichotomous:
  - OR=odds in the exposed/ odds in the unexposed

- If X is continuous:
  - OR=proportional increase in odds per one unit change in X.
  - Implies that odds increase by the same relative factor for each one unit change in X.
  - This implies the effect is multiplicative

Basic logistic model: Walk through examples of odds and probabilities for different values of X

- Logit(p)= ln(p/(1-p)) =$\alpha+\beta X$
- Logit(p)= ln(p/(1-p)) =$-3+.5*X$
  - OR=exp(.5)=1.65
- X=0,
  - logit=-3,
  - odds=exp(-3)=.05
  - p=exp(-3)/(1+exp(-3))=.05
- X=1
  - logit=-3+.5=-2.5
  - odds=exp(-2.5)=.082
  - .082/.05=1.64
  - p=exp(-2.5)/(1+exp(-2.5))=.075
  - .075/.05=1.5
- X=2
  - logit=-3+.5*2=-2
  - odds=exp(-2)=.135
  - .135/.05=2.7
  - p=exp(-2)/(1+exp(-2))=.119
  - .119/.05=2.38
- X=2 vs X=0
  - Exp(.5*2)=exp(1)=2.7

Basic logistic model, with two variables

- Logit(p)= ln(p/(1-p)) $=\alpha+\beta X$
- Logit(p)= ln(p/(1-p)) $=-3+.5*X+1.2*Z$
  - $OR_x=\exp(.5)=1.65$
  - $OR_z=\exp(1.2)=3.32$
- X=0, Z=0
  - logit=-3,
  - odds=exp(-3)=.05
  - p=exp(-3)/(1+exp(-3))=.05
- X=1, Z=1
  - logit=-3+.5+1.2=-1.3
  - odds=exp(-1.3)=.27
  - .27/.05=5.5
  - Exp(.5+1.2)=exp(1.7)=5.5
  - 1.65*3.32=5.5
  - p=exp(-1.3)/(1+exp(-1.3))=.214
  - .214/.05=4.3=PR

Basic logistic model

- Logit(p)= ln(p/(1-p)) =$\alpha+\beta X$
- This is a multiplicative model: the odds ratios multiply
- If someone is already really "high risk", a doubling of that risk (OR~2) is much greater in absolute terms than if someone is "low risk"
  - Risk of .01, doubled to .02, absolute change of .01
  - Risk of .2, doubled to .4, absolute change of .2
- If a population is very high risk, cutting that risk in half has much greater benefits in absolute terms (i.e. prevents more cases) than if a population is low risk.
- If you test interaction terms in logistic models, you are testing deviations from multiplicative relationships

How do interactions work in a logistic model?

- Logit(p)= ln(p/(1-p)) =−3+.5*X+1.2*Z+.2*X*Z
  - OR_x=exp(.5)=1.65 for Z=0
  - OR_z=exp(1.2)=3.32 for X=0
- X=0, Z=0
  - logit=-3,
  - odds=exp(-3)=.05
  - p=exp(-3)/(1+exp(-3))=.05
- X=1, Z=1
  - logit=-3+.5+1.2+0.2=-1.1
  - odds=exp(-1.1)=0.33
  - Odds for X=1, Z=1 / odds for X=0, Z=0= .33/.05=6.7
  - Could also calculate: Exp(.5+1.2+0.2)=exp(1.9)=6.7

Suppose Z is continuous

- Logit(p)= ln(p/(1-p)) =–3+.5*X+1.2*Z+.2*X*Z
  - OR_x=exp(.5)=1.65 for Z=0
  - OR_z=exp(1.2)=3.32 for X=0
- X=1, Z=0
  - logit=-3+.5
  - odds=exp(-2.5)=.082
  - p=exp(-2.5)/(1+exp(-2.5))=.076
- X=1, Z=1
  - logit=-3+.5+1.2+0.2=-1.1
  - odds=exp(-1.1)=0.33
  - Odds for X=1, Z=1 / odds for X=1, Z=0= .33/.082=4.02
  - Could also calculate: Exp(1.2+0.2)=exp(1.4)=4.06
- X=1, Z=2
  - logit=-3+.5+2*1.2+2*0.2=0.3
  - odds=exp(0.3)=1.35
  - Odds for X=1, Z=2 / odds for X=1, Z=0= 1.35/.082=16.46
  - Could also calculate: Exp(2*1.2+2*0.2)=exp(2.8)=16.4

Quantifying Interaction

P_{11}=Risk when X=1 & Z=1
P_{00}=Risk when X=0 & Z=0
P_{10}=Risk when X=1 & Z=0
P_{01}=Risk when X=0 & Z=0

Additive Interaction = $p_{11} - p_{10} - p_{01} + p_{00}$

Multiplicative Interaction = $\frac{p_{11}p_{00}}{p_{10}p_{01}}$

RERI = Relative Excess Risk due to Interaction

$= (p_{11}/p_{00}) - (p_{10}/p_{00}) - (p_{01}/p_{00}) + (p_{00}/p_{00})$

$= RR_{11} - RR_{10} - RR_{01} + 1$

Knol & VanderWeele recommended layout for interaction tables

Table 4 Example—interaction between XRCC1 codon 399 genotype and dietary vitamin E intake on the risk of prostate cancer

	XRCC1 codon 399 genotype				
	Arg/Gln + Gln/Gln		*Arg/Arg*		OR (95% CI) for *Arg/Arg* within strata of vitamin E
	N cases/controls	OR (95% CI)	N cases/controls	OR (95% CI)	
High vitamin E intake	17/46	1.0	14/44	1.04 (0.42–2.60); $P = 0.93$	1.04 (0.42–2.60); $P = 0.93$
Low vitamin E intake	22/57	1.22 (0.53–2.82); $P = 0.65$	24/33	2.40 (1.02–5.63); $P = 0.04$	1.97 (0.89–4.41); $P = 0.10$
ORs (95% CI) for vitamin E within strata of genotype		1.22 (0.53–2.82); $P = 0.65$		2.30 (0.93–5.74); $P = 0.07$	

Measure of interaction on additive scale: RERI (95% CI) = 1.14 (−0.57 to 2.85); $P = 0.19$.
Measure of interaction on multiplicative scale: ratio of ORs (95% CI) = 1.89 (0.56 to 6.34); $P = 0.30$.
ORs are adjusted for age, ethnicity, first-degree relative with prostate cancer, education, ever been a farmer, BMI, total energy intake, total fat intake and intake of other antioxidants.

- Show all relevant comparisons: P00 vs (P11, P01, and P10) and
- P11 vs P10
- P01 vs P00
- P10 vs P00
- P11 vs P01

Quantifying Interaction

Additive Interaction = $p_{11} - p_{10} - p_{01} + p_{00}$

RERI = Relative Excess Risk due to Interaction

$= (p_{11}/p_{00}) - (p_{10}/p_{00}) - (p_{01}/p_{00}) + (p_{00}/p_{00})$

$= RR_{11} - RR_{10} - RR_{01} + 1$

- RERI > 0 if and only if there is additive interaction > 0
- RERI > 0 implies the public health consequences of an intervention on X would be larger in the Z=1 group
- RERI < 0 implies the public health consequences of an intervention on X would be larger in the Z=0 group.
- Sufficient cause interaction means there are individuals for whom Y would occur if both X and Z were present but would not occur if just one or the other were present. Follows if RERI>1
- Epistatic interaction means individuals for whom Y would occur if and only if both X and Z are present. Follows if RERI>2

Uncertainty

Sources of uncertainty

- Is your code doing what you think it is doing?
- If you repeated this again with all exact same sampling procedure, but draw a new sample, and apply identical analysis procedures, what would be the distribution of this parameter?
- If I repeated this again, with identical data but left it to a reasonable scientist to make analytic decisions s/he thought best?
- If I repeated this again in a new setting with different populations, measurement instruments, scientists… would I find the same result?
- If I tried to test the theory in a whole new way, would I find the same result?

Sources of uncertainty

- Is your code doing what you think it is doing?
- If you repeated this again with all exact same sampling procedure, but draw a new sample, and apply identical analysis procedures, what would be the distribution of this parameter?
- If I repeated this again, with identical data but left it to a reasonable scientist to make analytic decisions s/he thought best?
- If I repeated this again in a new setting with different populations, measurement instruments, scientists… would I find the same result?
- If I tried to test the theory in a whole new way, would I find the same result?

Reproducibility

- Segregate data for projects
  - Raw
  - Recoded
- Separate from code
- Write all code in a script that ideally you can run top to bottom
  - Recode variables
  - Select analytic sample
  - Analyze data
- *Never* recode variables without renaming variable
- Document your code explaining what you are doing and why
- Try to line up a code reviewer- someone who can read your code, see if it corresponds with what you said in your methods, see if it produces what you expected (consider inviting this person to co-author)

Tips for writing code to be checked

- Provide an informative title to your code.
- Annotate your code. Well annotated code makes it easier for the code checker to know what you are trying to accomplish during each data/proc step (and makes it easier for them to tell if you accomplished what you think you have accomplished).
- Order your code. When submitting code to be checked, try to follow the order of the steps outlined in your methods and results section.
- Make sure the checker can tell where you pulled variables from. This is usually easy to do if you have libname statements, but if you used a more unusual dataset, it might be worth pointing this out to the code-checker.
- Clean your code. The code checking process will go faster if the coding of only relevant variables is included.
- Assume your code includes some errors, this is the reality of life.

Tips for checking code

- Look up variables you do not know. If the person is using a variable that you have not used before, it is worth looking it up in a codebook to determine if they pulled the correct variable and to find out the categories of the variable they are working with.
- Run the code yourself. By changing the libname directories, you can run the person's code and make sure that you get the same output they do.
- Program it another way. If the person is performing some data cleaning and you are unsure of how they cleaned the code, try programming it in a way you feel more comfortable with and see if you get the same answer.
- Check the paper against the code. Whatever the person says they did in the methods section should be in the code. Similarly, whatever is in the results needs to be in the code.
- Check the output. Run the tables/models yourself and check them again the results presented in the paper.
-

Sources of uncertainty

- Is your code doing what you think it is doing?
- If you repeated this again with all exact same sampling procedure, but draw a new sample, and apply identical analysis procedures, what would be the distribution of this parameter?
- If I repeated this again, with identical data but left it to a reasonable scientist to make analytic decisions s/he thought best?
- If I repeated this again in a new setting with different populations, measurement instruments, scientists… would I find the same result?
- If I tried to test the theory in a whole new way, would I find the same result?

Standard approaches to characterizing uncertainty

- Most emphasis on assessing uncertainty is on the possibility that the associations you observe in the sample you drew represent the associations in the population from which it was drawn.
  - Standard errors
  - Confidence intervals
  - Intuitions of bootstrapping
  - P-values
  - Evaluating subgroup effects

Uncertainty expressed as precision of estimates

- 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: $s/\sqrt{n}$
  - Variance of the estimator is the square of the standard error of the estimator: bigger SE implies bigger variance
- 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 $s/\sqrt{n}$

How to estimate the SE of the estimated mean of a variable?

Make up some data for a large sample N=10,000
X is a normally distributed random variable with a mean of 3.016 in the full sample

```
. set obs 10000
number of observations (_N) was 0, now 10,000

. gen id=_n

. gen x=rnormal(3,1)

. sum x
```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	10,000	3.015569	.9888733	-1.123581	6.69673

How to estimate the SE of the estimated mean of a variable?

Arbitrarily select 100 observations and estimate the mean

```
. mean x if id<101

Mean estimation
```

	Mean
x	2.968194

How confident am I in this estimate based on only 100 observations?

How to estimate the SE of the estimated mean of a variable?

- Use the SD and N to estimate the standard error

```
. mean x if id<101

Mean estimation                   Number of obs    =       100
```

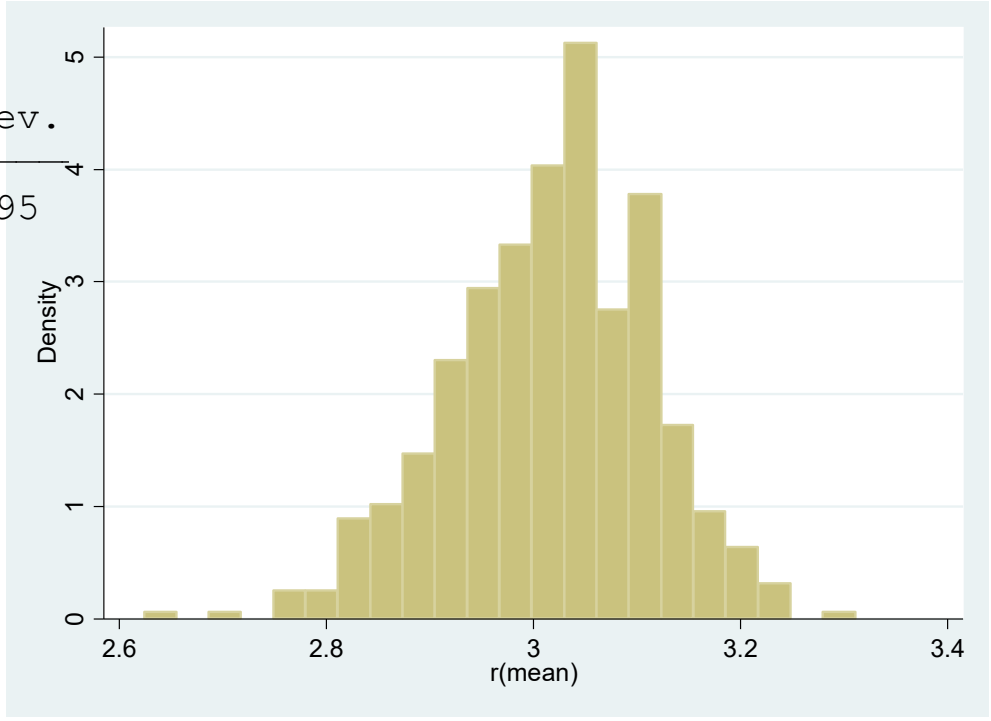
	Mean	Std. Err.	[95% Conf.	Interval]
x	2.968194	.0870845	2.795399	3.140988

How to estimate the SE of the estimated mean of a variable?

- Repeat the sample numerous times and observe the distribution of the estimated mean across each sample
- bootstrap r(mean) , si(100) sa(classse.dta, replace) r(500) :sum x

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
 sample_mean |       500    3.016679    .0963695   2.623821   3.310911
```



- Std devn across repeated samples approximates the std error calculated in 1 sample

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= $s/\sqrt{n}$, var=$s^2/(n)$
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)
4. Odds ratio: SE(ln(OR))= $\sqrt{\frac{1}{n_a}+\frac{1}{n_b}+\frac{1}{n_c}+\frac{1}{n_d}}$

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

The culture war around P-values

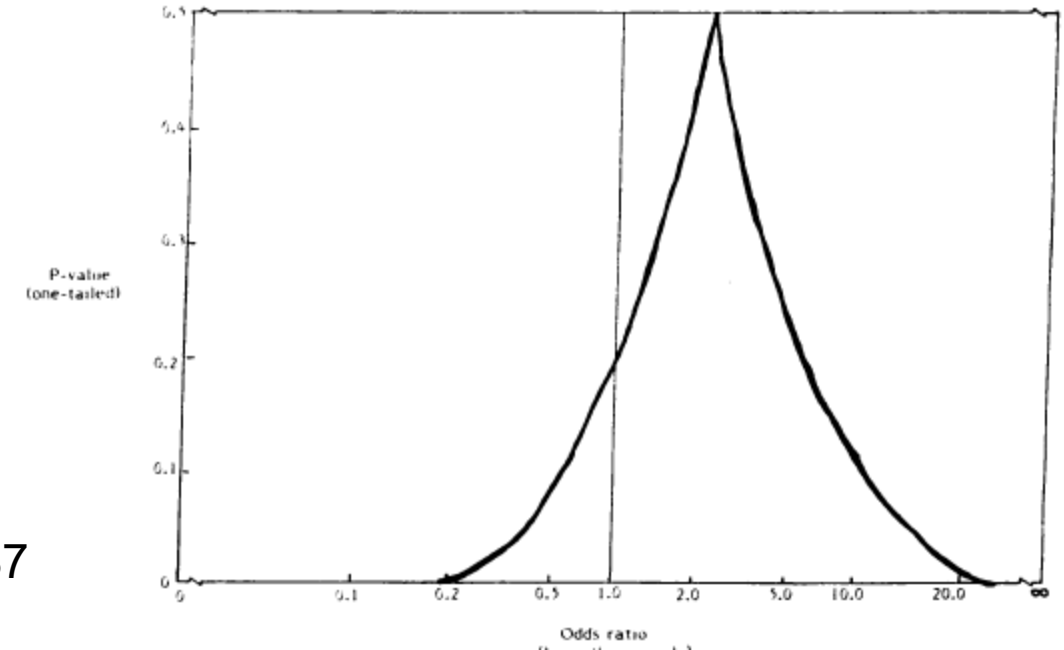
- Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value (ASA Statement, 2016)
- Usually the statistical model we are assessing is a null model of independence between two variables

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

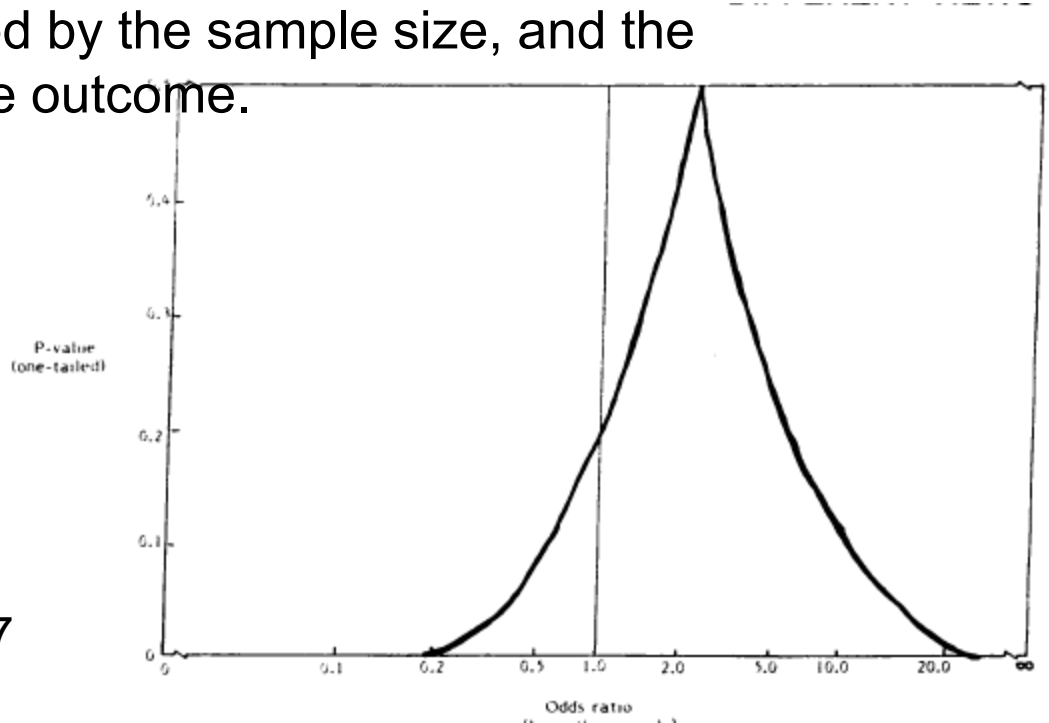
From: Poole, Beyond CIs, 1987



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

- 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?

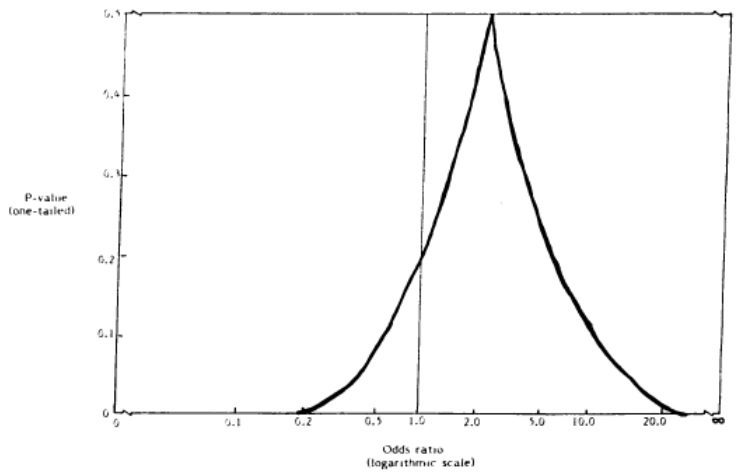
P-values: some common traps to avoid

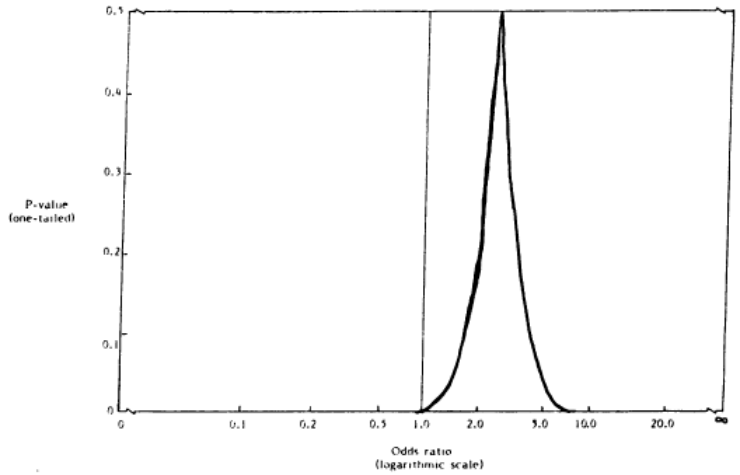
- Situations when this is tempting:
  - $E(BMI)=b_0+b_1*Mom\_Ed$ among women
  - b_1=-2 95% CI: (-.08, -3.92)
  - $E(BMI)=a_0+a_1*Mom\_Ed$ among men
  - a_1=-1.8 95% CI: (0.20, -3.80)

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

P-values: some common traps to avoid

- The p-value > .05, therefore, β=0





Conceptually, can plot the p-value for tests of whether data arose from any in a range of coefficients

P-value=1 when testing against the exact point estimate you found in your data. That's the coefficient *most consistent* with your data.

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

From: Poole, Beyond CIs, 1987

Confidence Intervals

- The set of parameter values for which the P-value exceeds some threshold, e.g., 0.05
- Calculate the P-value comparing the observed data to a particular "truth" (e.g., beta=0, 0.1, 0.2, 0.3…4.0, 4.1)
- Identify the range of possible betas for which the P-value is >0.05
- That's the 95% confidence interval for the beta
- A 95% CI will over unlimited repetitions contain the true parameter at least 95% of the time (if statistical model is correct and there's no bias)

Confidence Intervals

- A 95% CI will over unlimited repetitions contain the true parameter at least 95% of the time (if statistical model is correct and there's no bias)
- CI gives you more information on how much you've learned from the study compared to just a p-value
  - Very wide CI: not much info.
  - i.e., large SE and large variance
  - Sometimes show SE instead of CI
- Often the CI is just used as significance tests: if the CI doesn't include the null, P<.05 for test of the null. This is ignoring advantages of the CI.

P-value misuse rampant

- ASA issued a statement
- Many journals essentially ban p-values
- The use of P-value thresholds (* indicates P<.05) is even worse.
  - Encourages binary thinking: P<.05, so it's true!!!
  - Nobody can even back-calculate the SEs to use your results in future work

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:

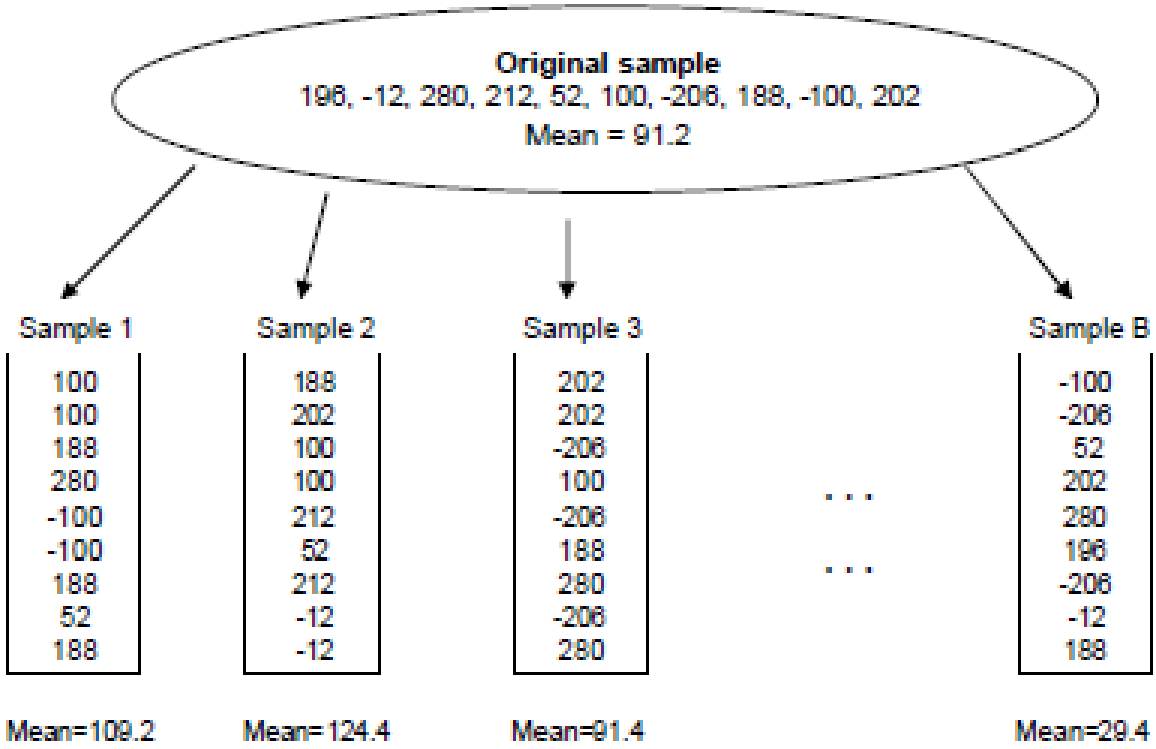
- The 25th percentile minus the 75th percentile
- The difference between two regression coefficients estimated from the same data in non-nested models.
  - Eg immediate risk vs cumulative risk models
- The mean predicted value for the top 20% of the distribution if you estimate among women only versus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

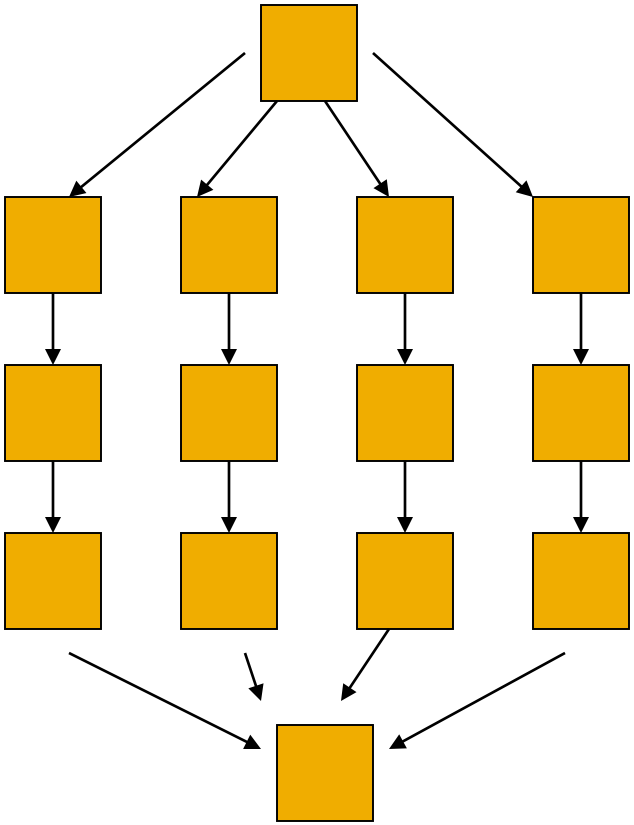
1. Estimate the missing value.
   - You now have several (e.g., 20) *possible* data sets
2. Estimate the parameter of interest (e.g your regression coefficient)
   - Derive one parameter estimate in each of the possible data sets.
   - Now you have a distribution of estimates, some bigger, some smaller
   - 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

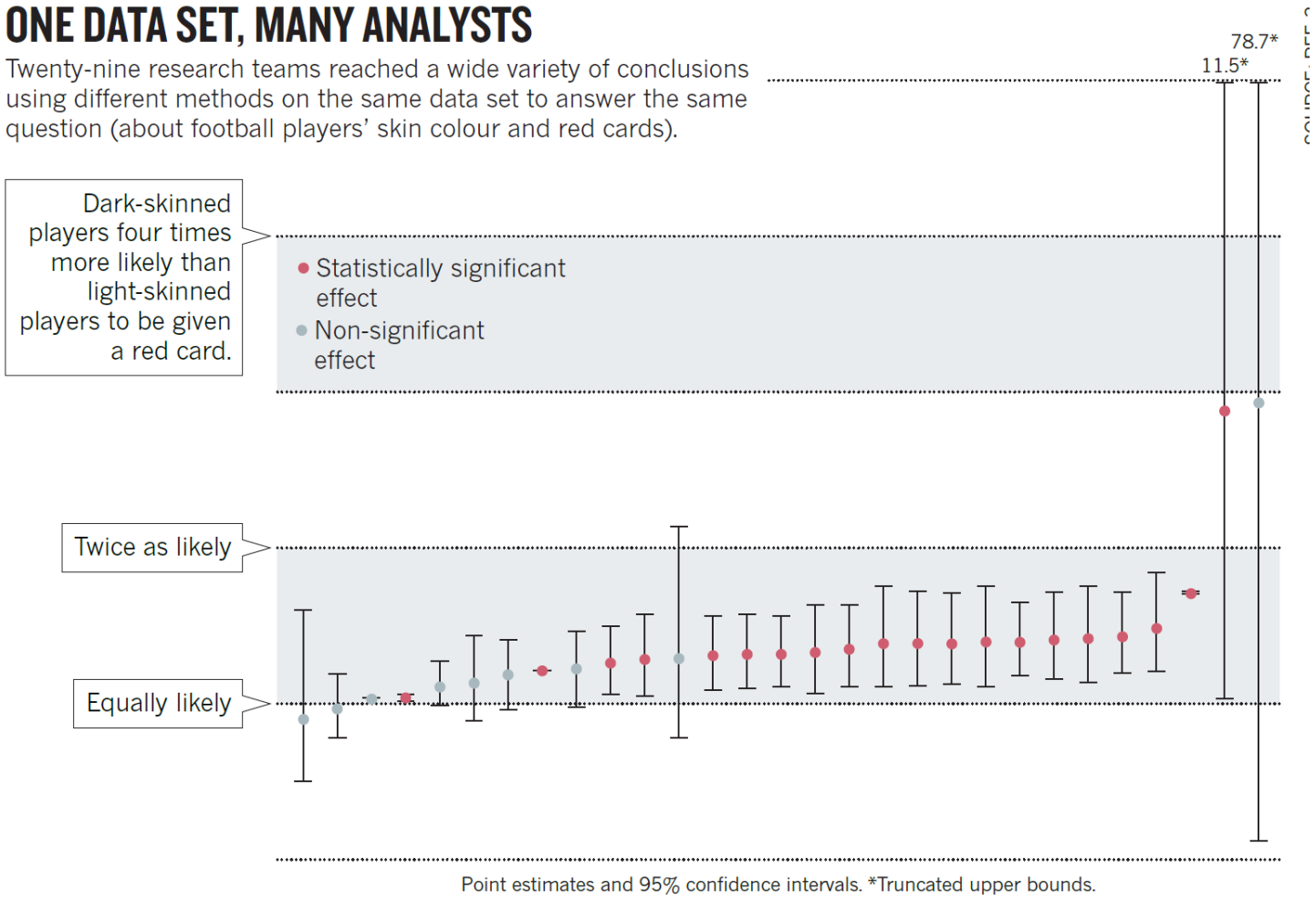
- This is automated in SAS and Stata
- Previously controversial because it is "making up data", increasingly standard
- Sometimes reviewers want to see it.
- If there is only a small fraction of missing data, it probably won't matter.
- If there's a lot of missing data, or **highly scatter-shot** missing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

Sources of uncertainty

- Is your code doing what you think it is doing?
- If you repeated this again with all exact same sampling procedure, but draw a new sample, and apply identical analysis procedures, what would be the distribution of this parameter?
- If I repeated this again, with identical data but left it to a reasonable scientist to make analytic decisions s/he thought best?
- If I repeated this again in a new setting with different populations, measurement instruments, scientists… would I find the same result?
- If I tried to test the theory in a whole new way, would I find the same result?

You stpped here!!!

Scientific judgment: critical but often ignored source of uncertainty



Silberzahn and Uhlmann 2015

Sources of uncertainty

- Is your code doing what you think it is doing?
- If you repeated this again with all exact same sampling procedure, but draw a new sample, and apply identical analysis procedures, what would be the distribution of this parameter?
- If I repeated this again, with identical data but left it to a reasonable scientist to make analytic decisions s/he thought best?
- If I repeated this again in a new setting with different populations, measurement instruments, scientists… would I find the same result?
- If I tried to test the theory in a whole new way, would I find the same result?

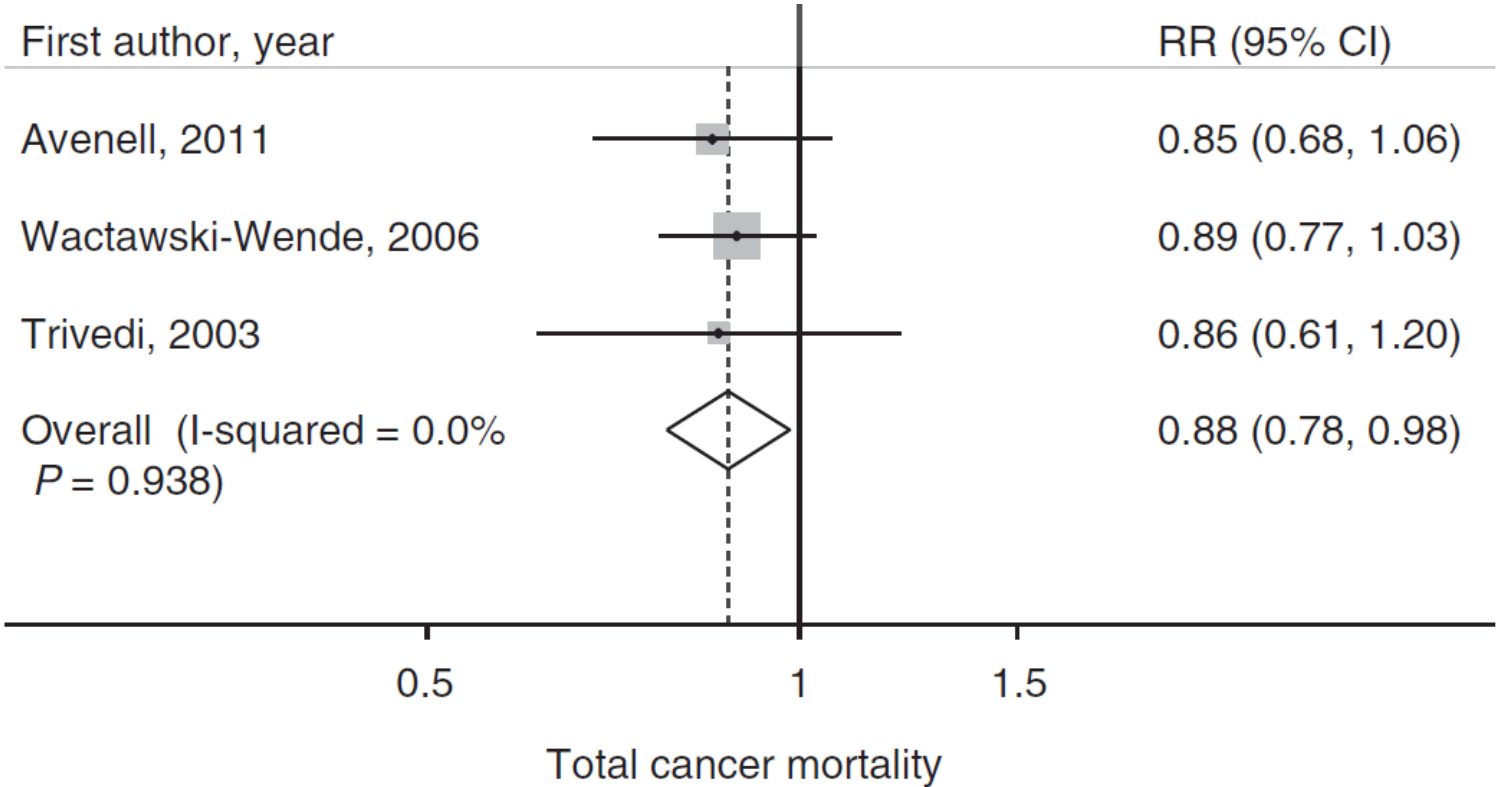
Meta-analyses

- Meta-analyses can help with three categories of uncertainty
  - Sampling variability because they effectively increase the sample size
  - Researcher judgment uncertainty to extent that researchers are independent
  - New setting uncertainty because often conducted with slight variations in population, measures, etc.

Meta-analysis of RCTs of vitamin D supplements and cancer mortality

Meta-analysis goal is to combine effect estimates usually weighting by the inverse of the variance (ie the information in the study).

No individual study convincing, but 3 studies, with nearly identical RRs, fairly convincing evidence



Meta-analysis: fixed effects model vs random effects?

- Use fixed effects if you believe the only reason effect estimates differ between studies is chance due to sampling variability for those enrolled in that study:
  - All studies are functionally identical (because study design would lead to differences in effect estimates
  - There is a single "true" effect size that prevails in all populations to which you are generalizing.
- Fixed effects model assumes the observed effect (β_i) for any study is:
  - $\beta_i=\theta+\varepsilon_i$
- Where θ is the true effect size, which is the truth for all study settings
- And ε_i is the difference between θ and the effect in this study, which is estimated by the standard error for the effect estimate in study *i*
- Each study is weighted (W_i) by the inverse of its variance: $V_i=\frac{\sigma^2}{n}$
  - Big variance→ imprecise study → count it less
- Calculate the pooled effect as $M=\frac{\sum_i^k W_i\beta_i}{\sum_i^k W_i}$

Standard Errors in the original study feed into the variance of the meta-analyzed effect estimate

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)
4. Odds ratio: SE(OR)= $\sqrt{\frac{1}{n_a}+\frac{1}{n_b}+\frac{1}{n_c}+\frac{1}{n_d}}$

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

Meta-analysis: fixed effects model vs random effects?

- Calculate the pooled effect as M= $\frac{\sum_i^k W_i \beta_i}{\sum_i^k W_i}$

Variance=$\sigma_\varepsilon^2/(n-1)S_X^2$
So, this is only the approximate relative value. To the extent that the variance of the exposure or the residual variance in Y differs between studies, this will not be right.

Study	N	b	Approximate Relative Variance (proportional to 1/N)	Weight (1/Variance)	% of total Weight	W*b
1	100	0.2	0.010	100	6%	20
2	1000	0.3	0.001	1000	64%	300
3	400	0.1	0.003	400	26%	40
4	64	0.6	0.016	64	4%	38.4
Sum	1564			1564		398.4
Pooled effect estimate						**0.25**

Standard Errors in the original study feed into the variance of the meta-analyzed effect estimate

Many statistics we use frequently are popular because they have known sampling distributions, so you can calculate their standa ver again

1. Me
2. Pro *q/(n−1)
3. Lin
   - s
   - (*the
     s
4. Od

With a fixed effects meta-analysis, the variance of the pooled effect estimate is the same as you would have gotten if you'd done one study that pooled all the individuals in your fixed effects meta-analysis.

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

Use random effects if effect estimates differ between studies due to both sampling variability within studies and differences in the true effect in each study.

- Random effects model assumes the observed effect (β_i) for any study is: $\beta_i=\mu+\zeta_i+\varepsilon_i$
- Where μ is the grand mean of effects across all studies
- ζ_i is the difference between the grand mean effect and the effect in study *i*
- ε_i is the sampling error for the effect estimate in study *i*, i.e., the difference between the true mean for study *i* and the observed mean for study *i*.
- Each study is weighted (W_i) by the inverse of *both* sources of variance: $W_i=\frac{1}{V_i+T^2}$
  - Study specific variance + Between study variance
- Calculate the pooled effect the same as for fixed effects: $M=\frac{\sum_i^k W_i\beta_i}{\sum_i^k W_i}$
- Trick is estimating the between-study variance (intuition: how much between study variance would be expect due to sampling? The rest is due to true effect heterogeneity)

Random, random everywhere…

- Remember our other uses of random effects models, where we used random deviations from an overall average to indexed clusters like a specific person (with repeated measures on that person) or a specific neighborhood (with multiple people observed within that neighborhood)?
  - Clusters with fewer observations got lower weight
  - We assumed that the overall average (random intercepts) or average effects (random slopes) differed between clusters
  - Best predictions for "truth" in any given cluster was a weighted average of the overall average and that particular cluster's values.
  - We asked questions about both the overall average effect estimate and about the variance between clusters.
- One other thing to remember for later (multiple imputation/Rubin's rules).
  - Each study is weighted (W_i) by the inverse of *both* sources of variance: $W_i=\frac{1}{V_i+T^2}$
  - Because the total uncertainty is the sum of: Study specific variance + Between study variance

Bayesian vs Frequentist Approaches

- Huge debate in biostatistics
  - From the outside, the Bayesians seem right, but the frequentists have the weight of tradition and simplicity of tools
- Most of intro stats, e.g., p-values and CIs, arise from a frequentist tradition, which assumes that parameters such as means, standard deviations, regression coefficients are fixed but unknown and estimated from a sample
  - We use stats to describe our expectations if we repeated our sampling procedure over and over, e.g., P-value for observed test statistic compared to a null hypothesis: frequency with which the test statistic would be this extreme or more so under the null hypothesis, assuming the model is correctly specified
- Bayesians assume:
  - We have a prior view about how likely any particular value for the parameters is (our priors: we would place a bet on any particular value, and we'd bet higher on some values than others).
  - When we receive new evidence, we should combine the new evidence with our priors to develop an updated view of the likely values for the parameters.
  - The updated view (the posterior probability), is more similar to the priors if we already had a strong bet, but more similar to the new evidence if we were very uncertain before collecting new data

Bayesian vs Frequentist Approaches, from Greenland slides:

- "Bayesian" is somewhat misleading: as with "Big Bang Theory", the label "Bayesian Statistics" came from a critic, R.A. Fisher, around 1950 (Fienberg, *BA* 2006).
- Unfortunately, that led teaching to focus on Bayes' theorem ($p(\theta|y) = p(y|\theta)p(\theta)/ p(y)$) for updating hypothesis probabilities.
- Other updating formulas (e.g., data augmentation; information addition) are more transparent and often ease computation as well.

external ("prior") information + internal study information = total (posterior) information

- Frequentist methods use no ***explicit*** prior, and so some claim the methods are "objective" or "let the data speak for themselves." This is pure delusion because frequentist methods are filled with ***implicit*** priors, and

DATA SAY NOTHING AT ALL!

- Data are markings on paper or bits that just sit there. **If you hear the data speaking, seek psychiatric care immediately!**

Bayesian approaches: key messages

- Combine prior information with new evidence
- You nearly always have a prior, either based on what is plausible or previous research
- Technical tools to articulate that prior and incorporate into analyses
- A practical approach: express priors as a small data set and combine with existing data before analyzing
- Many settings where a version of Bayesianism is used
  - Meta-analyses
  - Multi-level models
  - Likelihood ratios for evaluating diagnostic tests (sensitivity/(1-specificity))
  - Probability that conclusions are false (Ioannidis "Why most published research findings are false")

Sources of uncertainty

- Is your code doing what you think it is doing?
- If you repeated this again with all exact same sampling procedure, but draw a new sample, and apply identical analysis procedures, what would be the distribution of this parameter?
- If I repeated this again, with identical data but left it to a reasonable scientist to make analytic decisions s/he thought best?
- If I repeated this again in a new setting with different populations, measurement instruments, scientists… would I find the same result?
- If I tried to test the theory in a whole new way, would I find the same result?

Triangulation

- Strengthening causal inferences by integrating results from several different approaches, where each approach has different (ideally unrelated) key sources of potential bias – Lawlor (IJE 2016)
- Key sources of bias in multivariable regression in observational data (dominant approach in epi)
  - Unmeasured or poorly measured confounders
  - Reverse causation
  - Misclassification of exposure or outcome differential by the other
  - Differential missing data

Triangulation: Alternative Approaches

- Cross-context comparisons (new confounding structures)
  - Breast feeding in high versus low income countries
- Different control groups for case-control studies
- Natural experiments (new confounding structures, biases likely to differ)
- Within-person, within-family, within-twin studies
  - controls many types of confounders, exacerbates other biases
- IVs applied to RCTs
- Negative controls defined by theory
  - CSLs have no effect on health of college graduates

Ruling out alternative explanations with alternative study designs: example of Cancer & AD

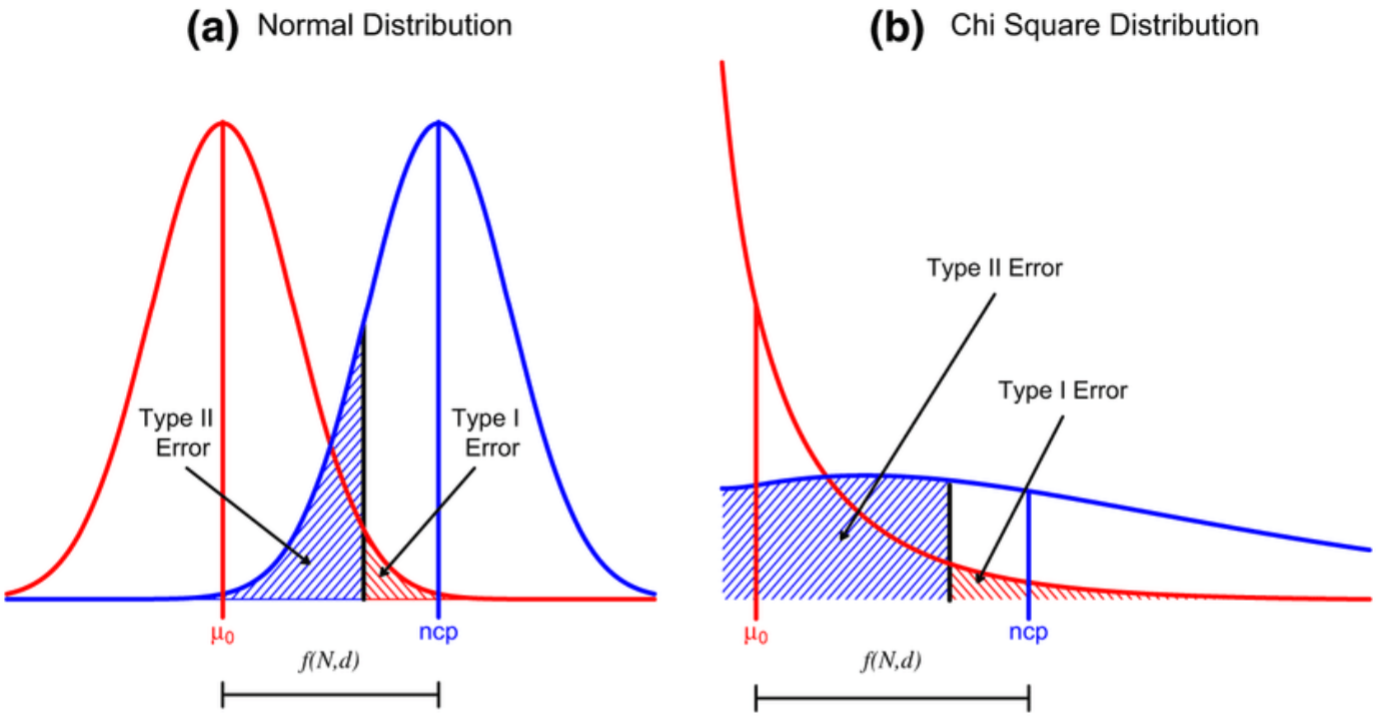
Testable implications of alternative explanations for the inverse comorbidity of cancer and AD		Consistent with:					
Experimental finding:	Evidence to Date?	Diagnostic Bias	Competing Risks	Survival Bias	Inverse Common Cause	Causality: Treatment	Causality: Physiologic Response
Lower lifetime risk of AD among cancer survivors	Strong	Yes	Yes	Yes	Yes	Yes	Yes
Lower incidence rate of AD among cancer survivors	Strong	Yes	No	Yes	Yes	Yes	Yes
Lower risk and rate of AD among people with non-lethal cancers	Mixed	No	No	No	Yes	Yes	Yes
Slower rate of cognitive decline after cancer diagnosis	Limited	No	No	Yes	Yes	Yes	Yes
Slower rate of cognitive decline before cancer diagnosis	None	No	Yes	Yes	Yes	No	Yes
Genetic variants related to cancer predict lower AD in people with no diagnosed cancer	None	No	No	No	Yes	No	Yes
Genetic variants related to AD predict lower cancer in people with no diagnosed AD	None	No	No	No	Yes	No	No
Genetic variants related to cancer predicts lower AD only in people with diagnosed cancer	None	Yes	Yes	Yes	No	Yes	Yes

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, 1 study design with potential biases implicit to that design.
- Seek replication and honest accounts of uncertainty
  - Uncertainty within studies
  - Uncertainty between studies
  - Both contribute to uncertainty

end

end



A graphical depiction of primary components of statistical power. The red distribution represents the distribution of test statistics under the null and the blue distribution represents the distribution of test statistics under the alternative. The black line indicates the significance threshold. The hatched red lines indicate the probably of a Type I Error or α Error. The hatched blue lines indicate the probably a Type II Error or β Error. Power is 1 − β. The non-centrality parameter (ncp), is the mean of the test statistic distribution under the alternative (Color figure online)