# Week 7: Supervised Class Measurement

## POLS0013 Measurement in Data Science

Dr. Julia de Romémont

Academic Year 24-25

UCL Departement of Political Science

# Last four lectures

## (Supervised) Scale Measurement

- Measures that summarise the *degree to which* the target concept is present/absent in the units of interest
- (Mainly)[1] interval-level scales where the relative ordering *and* distance between units of interest are *meaningful*
- We discussed various techniques that can be used to create such a scale by using previous knowledge (e.g. theory, expertise, training data)
  - Using (purely) theoretical reasoning (week 3)
  - Using theoretical reasoning to specify *linear* combinations of indicators (last week)
  - Using comparison data and (a form of) regression (week 4)
  - Using training data and (mainly) linear regression (week 5)

[1] Sometimes ratio-scales

# This lecture is about...

## (Supervised) Class Measurement

- Measures that denote *whether* the target concept is present/absent in the units of interest
- (Mainly)[2] nominal scales where the units of interest are assigned to *different classes*
- In this lecture we will discuss the techniques that can be used to create such a scale (e.g. theory, expertise, training data)
  - Using theoretical reasoning to create coding rules
  - Using training data and (logistic) regression

[2] Sometimes ordinal scales

# This Lecture

Classifying stuff

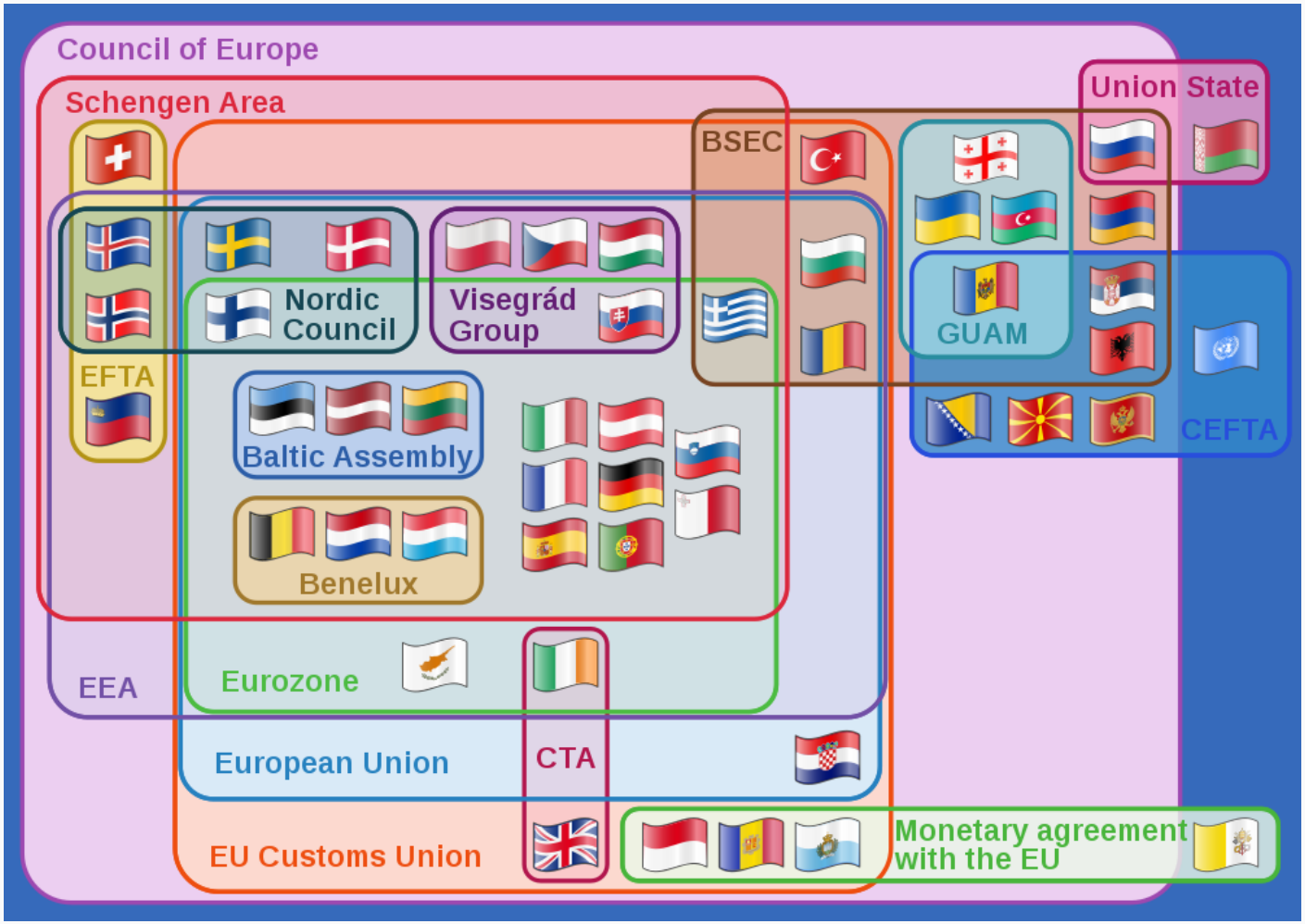
...Using coding rules

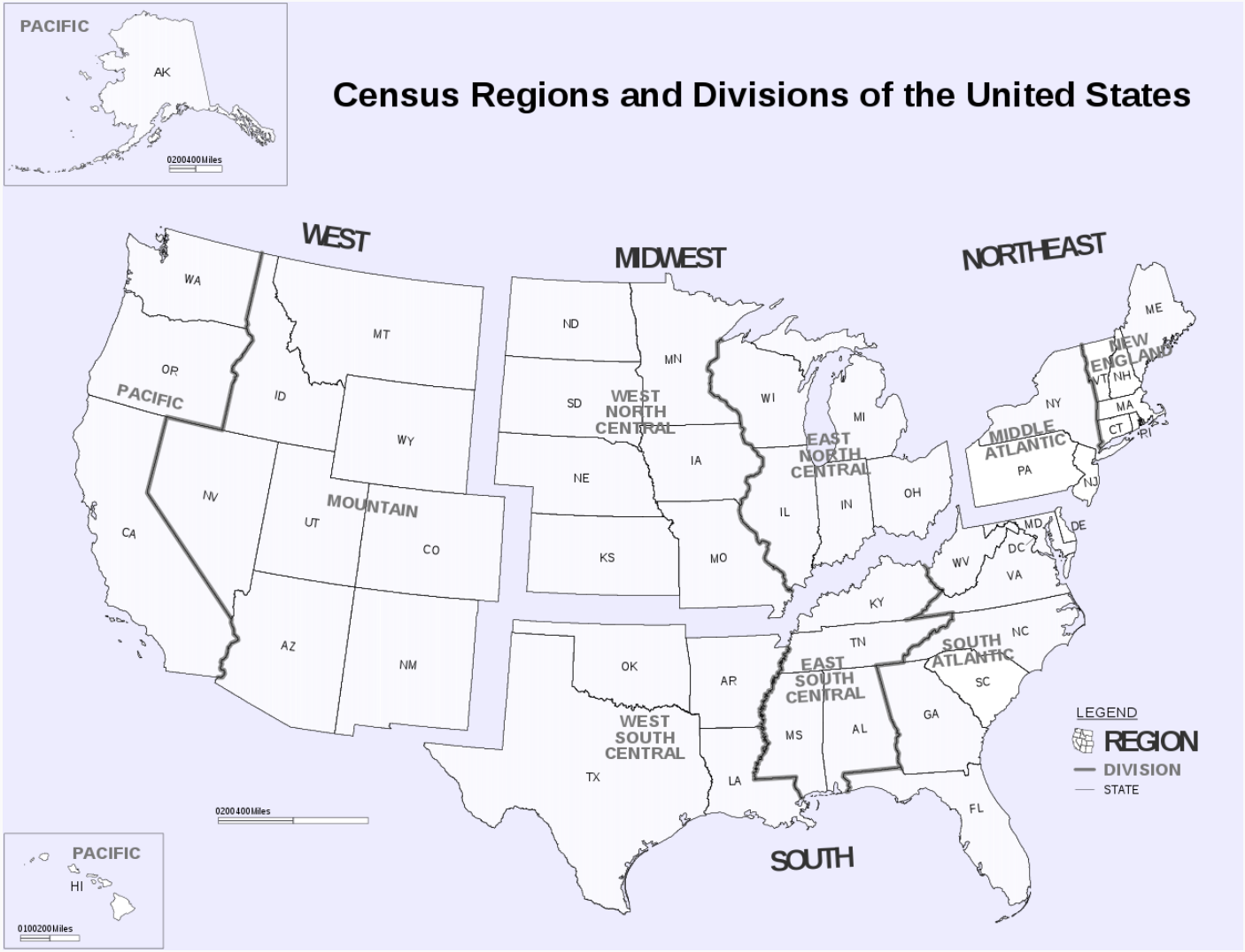
...Using training data

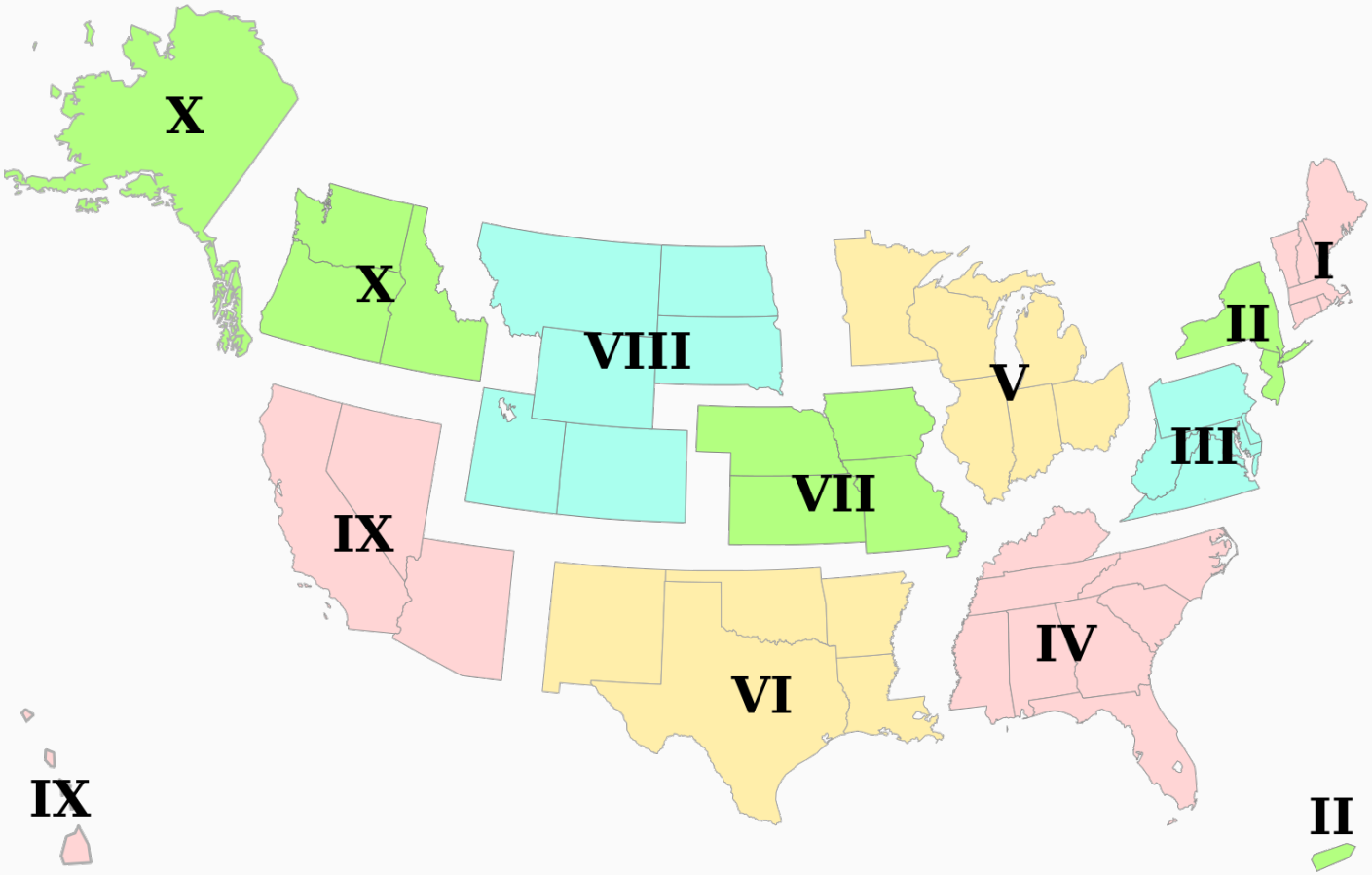
Measurement error in nominal scales

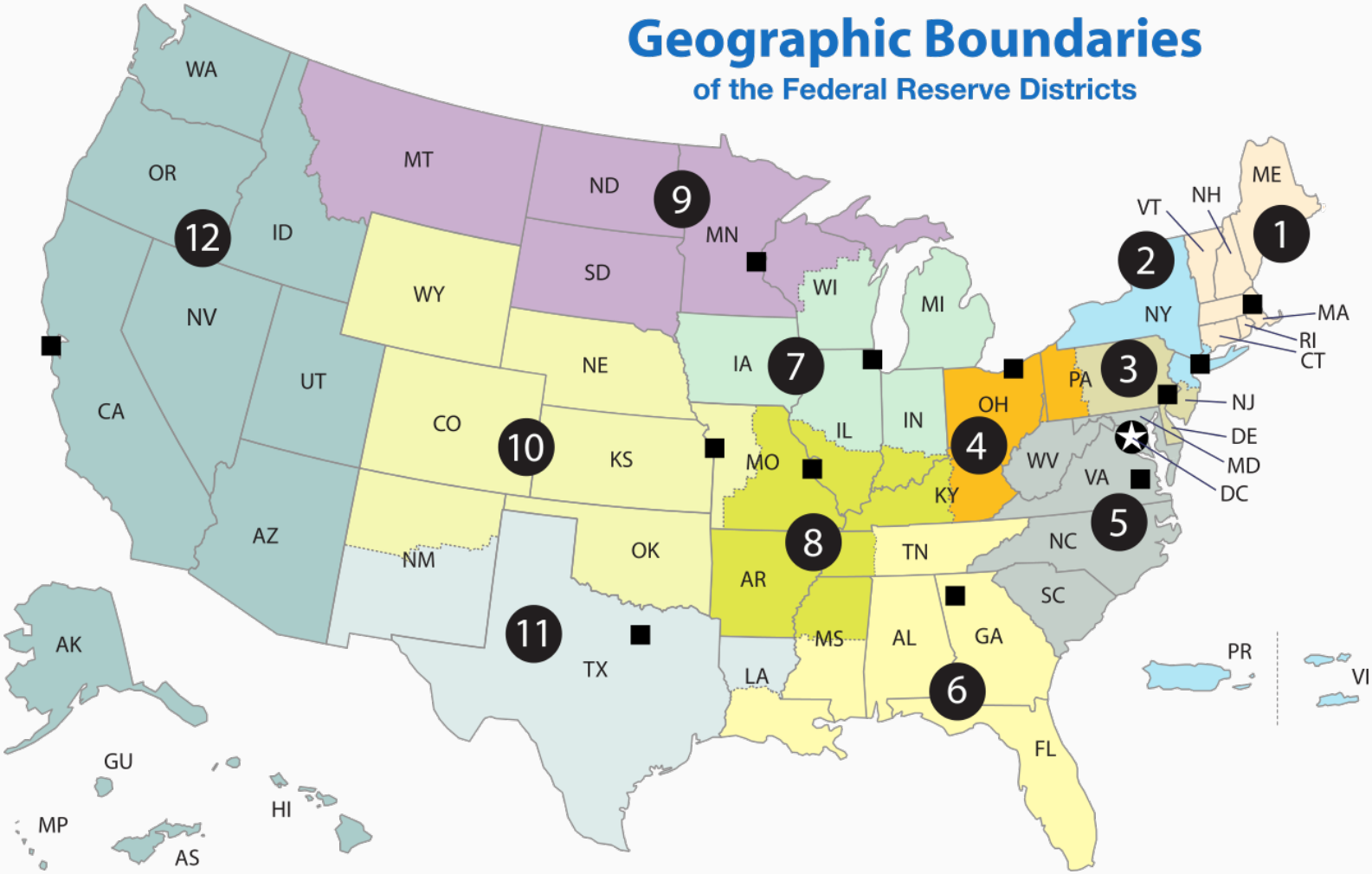
Point vs probabilistic classifications

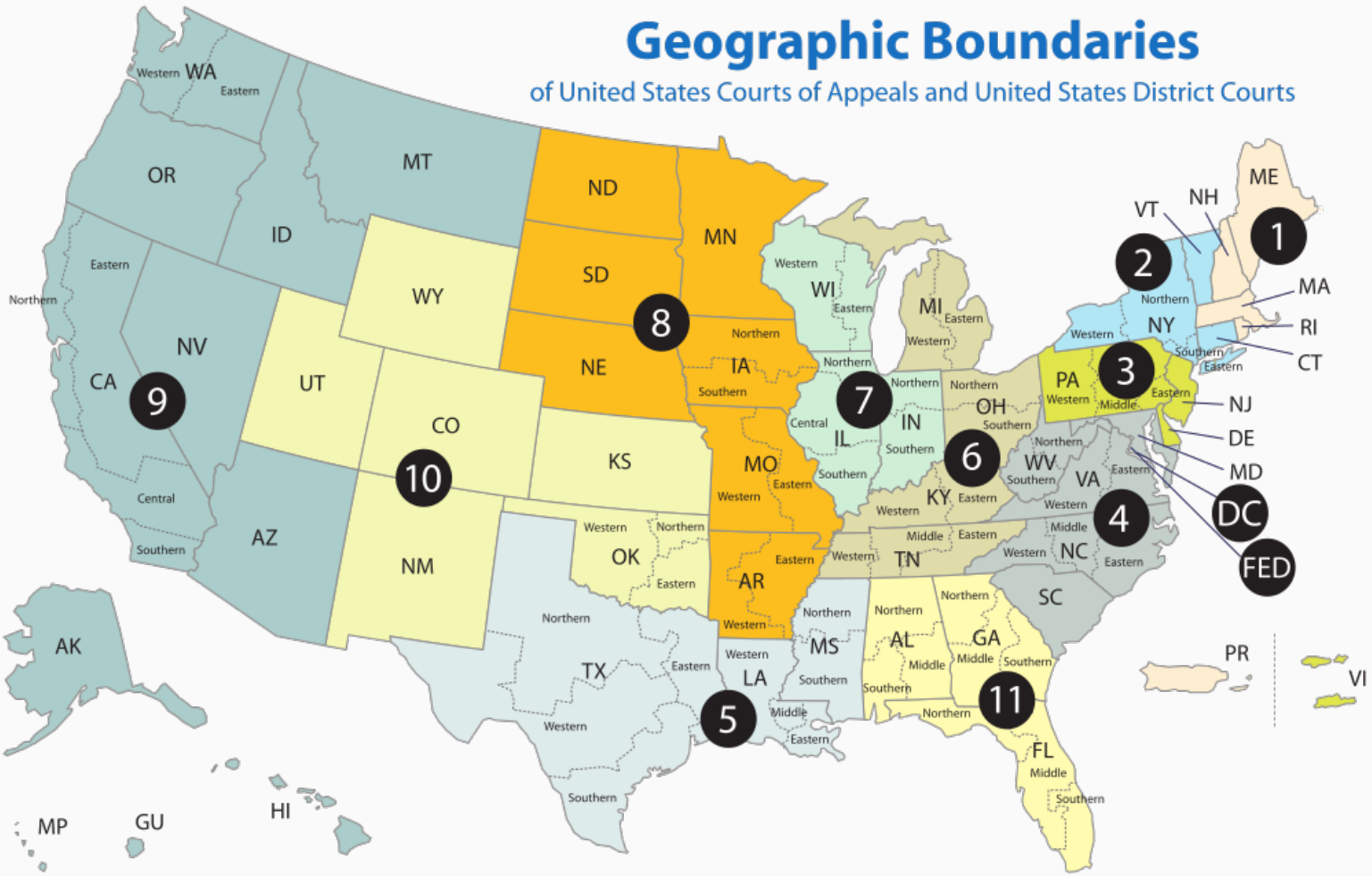
# Classifying stuff

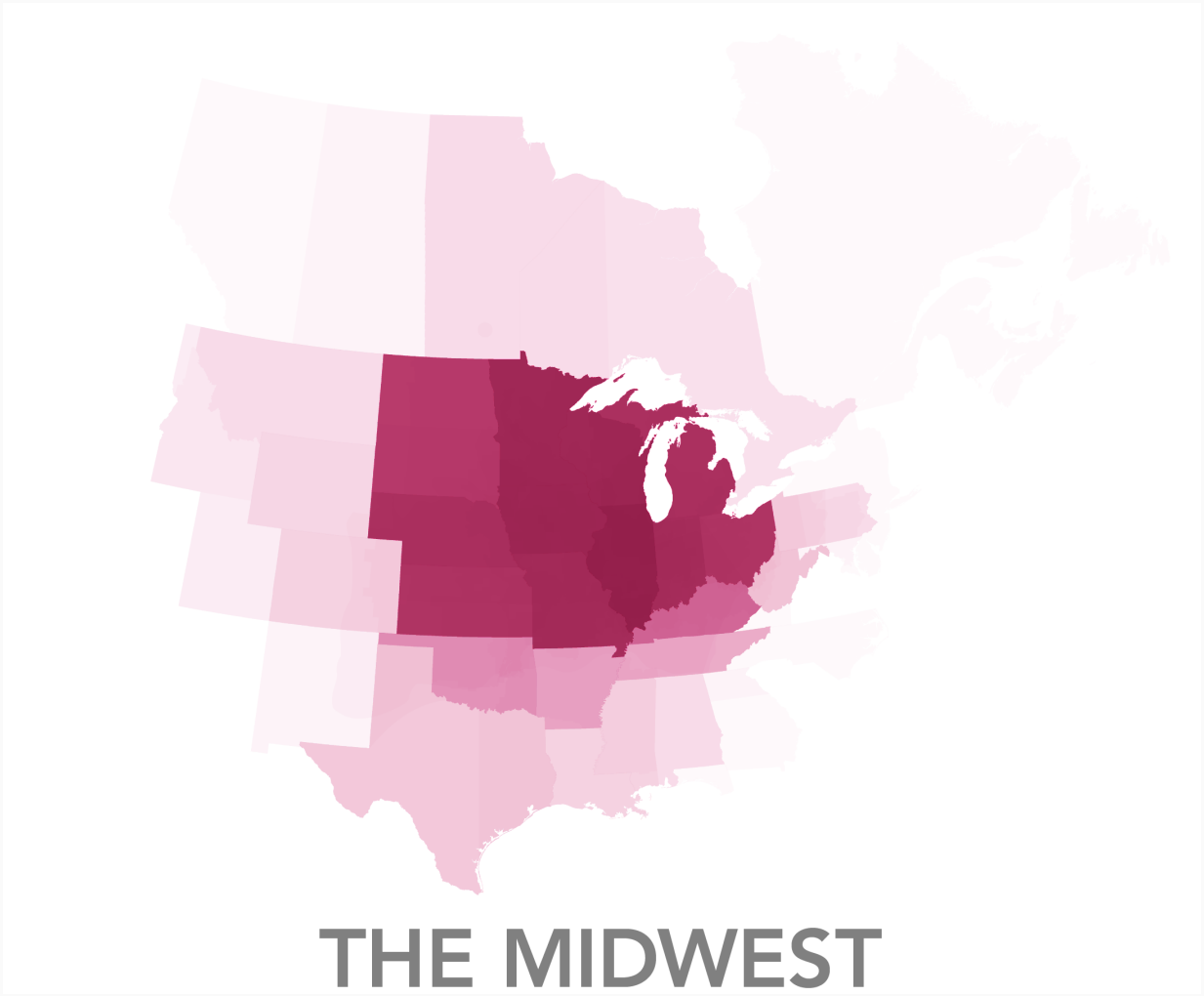

Council of Europe
Schengen Area
Union State
BSEC
Nordic Council
Visegrád Group
GUAM
EFTA
Baltic Assembly
CEFTA
Benelux
EEA
Eurozone
European Union
CTA
EU Customs Union
Monetary agreement with the EU


Census Regions and Divisions of the United States
PACIFIC
AK
0200400Miles
WEST
MIDWEST
NORTHEAST
SOUTH
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
MN
IA
MO
WI
IL
MI
IN
OH
KY
TN
MS
AL
OK
AR
TX
LA
WV
VA
NC
SC
GA
FL
PA
NY
ME
VT
NH
MA
CT
RI
NJ
MD
DE
DC
WEST NORTH CENTRAL
EAST NORTH CENTRAL
MOUNTAIN
PACIFIC
NEW ENGLAND
MIDDLE ATLANTIC
EAST SOUTH CENTRAL
WEST SOUTH CENTRAL
SOUTH ATLANTIC
LEGEND
REGION
DIVISION
STATE
0200400Miles
PACIFIC
HI
0100200Miles


X
X
VIII
V
I
II
III
VII
IX
IV
VI
IX
II


Geographic Boundaries
of the Federal Reserve Districts
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
PA
VA
NC
SC
NY
VT
NH
ME
MA
RI
CT
NJ
DE
MD
DC
PR
VI
AK
HI
GU
MP
AS
9
12
7
10
8
11
4
3
2
1
5
6


Geographic Boundaries
of United States Courts of Appeals and United States District Courts
Western WA
Eastern
OR
ID
MT
ND
SD
MN
WY
NE
8
Northern
IA
Southern
Eastern
NV
CA
9
Northern
UT
CO
10
KS
MO
Western
Eastern
Central
Southern
AZ
NM
Western
Northern
OK
Eastern
AR
Eastern
Western
Northern
Western
WI
Eastern
MI
Eastern
Western
7
Northern
IL
Central
Southern
IN
Northern
Southern
6
OH
Northern
Southern
KY
Eastern
Western
TN
Middle
Eastern
Western
VA
Eastern
Western
WV
Northern
Southern
NC
Middle
Western
Eastern
4
SC
Northern
Middle
Southern
GA
Northern
Middle
Southern
AL
Northern
Middle
Southern
MS
Northern
Southern
LA
Western
Middle
Eastern
5
TX
Northern
Western
Southern
Eastern
11
Northern
FL
Middle
Southern
PA
Western
Middle
Eastern
3
NY
Western
Northern
Southern
Eastern
2
VT
NH
ME
1
MA
RI
CT
NJ
DE
MD
DC
FED
AK
HI
MP
GU
PR
VI




A consensus definition of “The Midwest” based on 100 classifications by different organisations.

# Classification

- Measuring classes (categorical variables) rather than scales (continuous variables) is a choice of the analyst
- Some concepts can plausibly be measured either way:
  - Continuous scale for democracy: to what extent are countries democratic?
  - Categorical classification for democracy: is this country a democracy or not?
- Classification requires making sharp choices at the margins
  - Sometimes this is what you want to do, sometimes it is not.
  - Most of the time, where exactly the cut-off point is feels somewhat arbirtrary

# Making sharp choices has consequences

- In 2023, the US Census Bureau proposed to change how to measure whether a respondent is classified as disabled
  - This sparked a lot of (political) debate... and was eventually scrapped
- Old version: *"[...] six yes-or-no questions—related to difficulty with hearing, vision, and other functions—to determine disability status. A respondent who answers "yes" to any of those questions is counted as disabled.*
- Proposed change: *"six questions that cover the same topics as the ACS questions. Instead of answering "yes" or "no," however, respondents rate their level of difficulty on each function by choosing one of four options ranging from "no difficulty" to "cannot do at all."*
  - Cut-off at "a lot of difficulty"
- Have a think: what are the trade-off between these two different ways of classifying individuals as disabled?

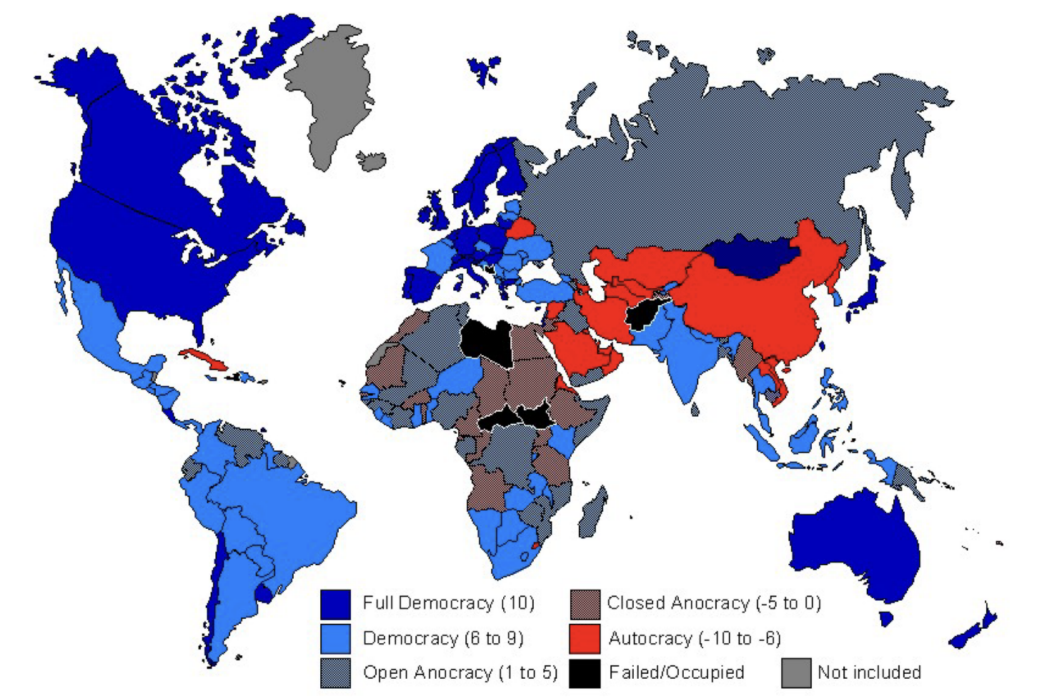
# ...Using coding rules

# Same old, same new[3]

- The intuition here is the same than what we had last week
- The measures are constructed by
  1. Conceptualisation
  2. Measurement
  3. Aggregation
- ... only that the target concept is nominal (or ordinal) and therefore the aggregation should reflect that at some stage, either
  - at the very end, by recoding a scale measurement, or
  - during, by recoding indicators and how they aggregate

[3] New, because we talked about it just last week

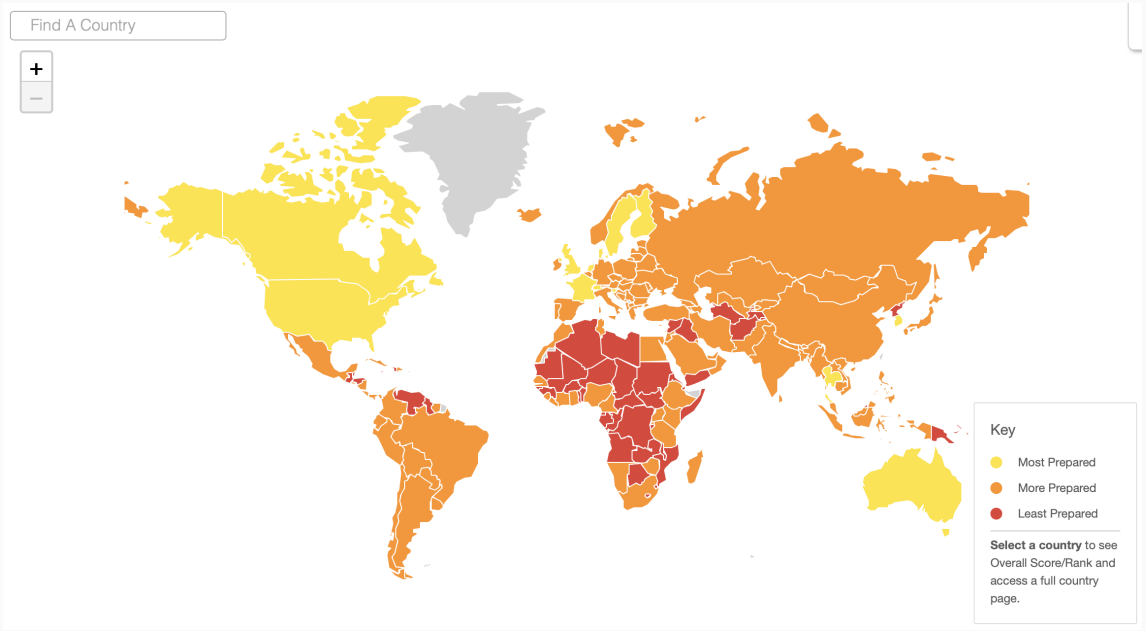
# Making categorical measures out of scales




Polity IV

- Some categorical measures are simply the result of recoding a continuous measures
- How useful do you think this is here? What are the advantages and disadvantages?

# Making categorical measures out of scales




GHS Index

- How meaningful are the categories here?
- Unless you have a good argument that the thresholds are meaningful, these are just throwing information away!

# Making categorical measures by recoding indicators

- Some categorical variables can be derived from existing categorical indicators using a set of coding rules
- This implies that the mapping of indicators to the variable of interest is specified by the analyst, for instance whether
  - All indicators are *necessary*[4]
  - Only one indicator is *sufficient*[5]
  - Or something in between, where only some of the criteria must be fulfilled
- This is analogous to index construction, but for a categorical measure rather than an interval level measure

[4] Implies a multiplicative aggregation

[5] Like in the example on slide 12.

# Example: binary coding scheme for democracy

- Binary classification of regimes into democracies and non-democracies by Alvarez et al (1996)
- Four rules for classifying which countries were non-democracies in which years.
  - Rule 1. "Executive Selection." The Chief Executive is not elected.
  - Rule 2. "Legislative Selection." The Legislature is not elected
  - Rule 3: "Party." There is no more than one party.
  - Rule 4: *A complicated rule about observed transitions of power through elections.*
- Any one of these rules applying to a country-year is **sufficient** to classify that country-year *non-democratic*.
  - The aggregation rule is multiplicative (all of the above are required) rather than additive.

# Content analysis

- **Content analysis**[6] can be classified[7] as an example of categorical measurement
  - Texts, audio, video, etc are *classified* or *coded* as containing (or failing to contain) certain "content"
  - The assessment is *qualitative* (by a human, reading)
  - Texts are classified either as a whole or, more typically, in short subsections.
  - Analysis of patterns of these *codes* or *classifications* can proceed quantitatively or qualitatively.
- Calling something "content analysis" usually implies additional things about what these codes are and how they are organised (outside our scope here)

[6] A prominent method used for analysis of documents.

[7] Pun intended.

# Deduction vs Induction

## Deductive Reasoning

Theory $\longrightarrow$ Predictions $\longrightarrow$ Conclusion about observations

- Content analysis sometimes proceeds deductively
  - A set of categories are determined for the purposes of a research question, and then documents are coded following the theoretically determined coding rules

## Inductive Reasoning

Observations $\longrightarrow$ Generalisation $\longrightarrow$ Theory

- Content analysis sometimes proceeds inductively
  - A set of categories is developed to describe relevant variation as the documents are assessed

# Example: Comparative Manifesto Project

- One prominent political science content analysis project is the Comparative Manifesto Project
- Each sentence of each party manifesto is coded into a very large number of categories based on a detailed set of coding rules
- Lots of challenges associated with doing this!

*"The Manifesto Project developed a category system whereby each quasi-sentence of every manifesto is coded into one, and only one, of 56 standard categories. The 56 categories are grouped into seven major policy areas and are designed to be comparable between parties, countries, elections, and across time."*

# Example: Coding a sentence

***1) Ambiguity of Language***

a) Often, political actors make policy statements by mentioning a negative aspect of an issue in order to highlight its importance. Take, for example, the following:

> *"Our country's democracy does not work well enough anymore!"*

This sentence could be read and interpreted as a negative statement towards the country's democratic processes. However, it is rather clear that the party is not making a statement against democracy itself. The actual message of this sentence is one of concern about and criticism of the current state of democracy. Therefore, this is a positive statement towards the ideal principle of democracy.

# Classifying stuff using coding rules

- It is difficult to say much about coding rules in general
  - They are meant to *encode* substantive expertise that is domain specific
- Stronger supervision (=more expertise!) is usually better
  - Where it is possible to carefully *directly* link the categorical quantity that you want to measure to already observed indicators, that is a good thing and you should do it!
  - Failing this, if you can carefully specify how a human coding of cases should proceed and how a scale should be split into categories, you should do this.
- Except when it isn't
  - If you want to summarise variation rather than measure a specific quantity of interest, *inductive* human coding may be appropriate.

# ...Using training data

# Making categorical measures with training data

- We previously considered running regression models predicting "gold standard" measures using a set of indicators.
- We could then use those same indicators, measured for a larger set of units, to construct measures (i.e. fitted values) using the estimated regression equations
- We can apply the same logic but using a limited dependent variable regression model.
  - Classification into *two* categories $\longrightarrow$ use binary logistic regression
  - Classification into *more than two ordered* categories $\longrightarrow$ use ordinal logistic regression
  - Classification into *more than two unordered* categories $\longrightarrow$ use multinomial logistic regression.

# Same challenges as with interval-level scales

- When we were considering the measurement of interval-level quantities, we worried about
    1. the *availability* of a training set with any pre-existing measures and relevant indicator data
    2. the *quality* of those pre-existing measures and of the indicators in the training set
    3. the *relevance* of those pre-existing measures and their relationship to the indicators to the target population
- All of these are still key issues in the categorical case

# Logistic regression, again

- There are many methods we could use to train a classifier
  - Again, the machine learning literature is full of possibilities
- Logistic regression solves the problem pretty well, most of the time
  - Regularization may be helpful, if you have many candidate indicators
- All these methods are tools for estimating the relationship between the indicator values you can observe and the classes/categories that you want to measure

# Training with logistic regression

- Training data $m \in 0, 1$ from some pre-existing measurement procedure for the concept of interest $\mu \in 0, 1$.
- One or more indicators $I$ ($I_1$, $I_2$, etc) that we want to use to measure the concept of interest.
- Remember, logistic regression can be expressed as the **log-odds ratio** of the probability that a variable (here: our measure $m$) is one $p(m_i = 1)$

$$\log \frac{p(m_i = 1)}{p(m_i = 0)} = \alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots$$

# Point classifications and probabilistic classifications

The predicted probabilities of a logistic regression are given by the formula:

$$\widehat{p(m_i = 1)} = \frac{e^{\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots}}{1 + e^{\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots}}$$

**What is our measurement?** It could be either:

1. The **predicted probability** $\widehat{p(m_i = 1)}$
2. A **point prediction** for $m_i$, where $\widehat{m_i} = 0$ if $\widehat{p(m_i = 1)} < 0.5$ and $\widehat{m_i} = 1$ if $\widehat{p(m_i = 1)} \geq 0.5$.

# Point classifications or probabilistic classifications?

- The point classification $\widehat{m_i} \in 0, 1$ is binary, like the target concept
- The probabilistic classification $p(\widehat{m_i = 1}) \in [0, 1]$ incorporates uncertainty about the true classification of the unit
- To think about which of these is actually what we want, let's take a brief detour to think about measurement error in binary variables...

# Measurement error in nominal scales

# Measurement errors in categorical variables

- The structure of errors is limited for a categorical variables
- With binary quantities $\mu \in 0, 1$...
  - ...for either true value of $\mu$ there is one correct value of $m$ and one incorrect value of $m$.
  - ...if $\mu_i = 0$, then $m_i = 1$ is a *false positive*
  - ...if $\mu_i = 1$, then $m_i = 0$ is a *false negative*.
- Bias and variance are not good ways of describing the ways in which binary variables can be wrong

# Still not simple

- There are a very large number of ways to describe errors in binary variables.
- The rate of errors when $\mu = 0$ can be very different than the rate of errors when $\mu = 1$, because one of these may be far more common than the other.
- One of these errors may be far more important than the other depending on the substantive context as well.

# Key terminology

| | $m = 0$ | $m = 1$ |
|---|---|---|
| $\mu = 0$ | true negative | false positive |
| $\mu = 1$ | false negative | true positive |

The above 2x2 matrix of possibilities is often called the *confusion matrix*.

# Accuracy & Total Error Rate

## Accuracy (Binary Variables)

$=$ the proportion of cases in which $m = \mu$

$$\frac{p(\text{true positive}) + p(\text{true negative})}{p(\text{true positive}) + p(\text{true negative}) + p(\text{false positive}) + p(\text{false negative})}$$

## Total Error Rate

$=$ the proportion of cases in which $m \neq \mu$

$$\frac{p(\text{false positive}) + p(\text{false negative})}{p(\text{true positive}) + p(\text{true negative}) + p(\text{false positive}) + p(\text{false negative})}$$

- The best possible values for each of these are 1 for the former and 0 for the latter

# Sensitivity & Specificity

## Sensitivity (Se)

$=$ the **true positive rate**

$$p(m = 1|\mu = 1) = \frac{p(\text{true positive})}{p(\text{true positive}) + p(\text{false negative})}$$

- i.e. the proportion of the cases where $\mu = 1$ for which $m = 1$
- i.e. the rate at which the model/measure correctly 'catches' a true value of 1

## Specificity (Sp)

$=$ the **true negative rate**

$$p(m = 0|\mu = 0) = \frac{p(\text{true negative})}{p(\text{true negative}) + p(\text{false positive})}$$

- i.e. the proportion of the cases where $\mu = 0$ for which $m = 0$
- i.e. the rate at which the model/measure correctly 'catches' a true value of 0

# Sensitivity & Specificity

- The best possible value for both of these is 1
- You can always achieve this for either sensitivity or specificity at the expense of the other
  - Setting $m = 1$ for all units (achieving perfect sensitivity)
  - Setting $m = 0$ for all units (achieving perfect specificity)
- Sensitivity and specificity **condition** on the true value $\mu$
  - What is the proportion of correctly measured values $m$ among units with a given true value $\mu$?

# Positive & Negative predictive value

- If we condition on $m$ instead of $\mu$, we get...

## Positive Predictive Value (PPV)

$$p(\mu = 1|m = 1) = \frac{p(\text{true positive})}{p(\text{true positive}) + p(\text{false positive})}$$

- i.e. the proportion of the cases where $m = 1$ for which $\mu = 1$
- i.e. the rate at which a predicted value of 1 by the the model/measure is truly a 1

## Negative Predictive Value (NPV)

$$p(\mu = 0|m = 0) = \frac{p(\text{true negative})}{p(\text{true negative}) + p(\text{false negative})}$$

- i.e. the proportion of the cases where $m = 0$ for which $\mu = 0$
- i.e. the rate at which a predicted value of 0 by the the model/measure is truly a 0

# Point vs probabilistic classifications

# Point vs probabilistic classification

- Recall, if we train a model, our measurement could either be:
  1. The predicted probability $\widehat{p(m_i = 1)}$
  2. A point prediction for $m_i$, where $\widehat{m_i} = 0$ if $\widehat{p(m_i = 1)} < 0.5$ and $\widehat{m_i} = 1$ if $\widehat{p(m_i = 1)} \geq 0.5$.[8]
- It seems like 2 is the right choice given that we are imagining the quantity of interest as binary, however...
- If you use point classifications, the bias in your mismeasured binary variable is almost *guaranteed* to be correlated with any other variable that the true values of that binary quantity were correlated with.
- This in turn almost ***guarantees*** **biases in any analysis that you might use these measures for**

[8] Or any other threshold value, although 0.5 often makes the most sense.

# Example: COVID diagnostic tests

Hypothetical example: you are assessing some possible long-run consequence of a person contracting COVID-19

- You are interested in whether some outcome $Y$ is different among those who caught COVID ($\mu = 1$) and those who did not catch COVID ($\mu = 0$).
- You have something like an antibody test $m \in 0, 1$ which is an imperfect measure of whether $\mu = 0$ or $\mu = 1$. It has:
  - A specificity of 100% ($Sp = 1$), i.e. correctly identifies *all* those who *didn't* have COVID.
  - A sensitivity of 90% ($Se = 0.9$), i.e. correctly identifies *90%* of those who *did* have COVID.

# Example: COVID diagnostic tests

- The true difference in means between those who had COVID-19 and those who did not is:

$$\Delta = E[Y|\mu = 1] - E[Y|\mu = 0]$$

- However, we only know the test result $m$, not the reality $\mu$, for each observed individual.
- Let's assume that the expected value of $Y$ depends only on $\mu$, not $m$:

$$E[Y|\mu, m] = E[Y|\mu]$$

  - This means that the expected value of $Y$ does not depend on the measurement error
  - This may or may not be reasonable for a given outcome, but is a best case.

# Consequences of ignoring the problem

All of those with $m = 1$ also have $\mu = 1$

- Because specificity of 100% ($Sp = p(m = 0|\mu = 0) = 1$), there are **no false positives**
- Since we assumed $E[Y|\mu, m] = E[Y|\mu]$...
- ...the mean of $Y$ among those who test positive is an unbiased estimate for everyone who had COVID-19:

$$E[Y|m = 1] = E[Y|\mu = 1]$$

# Consequences of ignoring the problem

However, not all of those with $m = 0$ also have $\mu = 0$

- Where the true proportion of those with COVID-19 is $\pi = p(\mu = 1)$...
- ...the proportion of those with negative tests who actually had COVID ($p(\mu = 1|m = 0)$[9]) is $\frac{(1-Se)\cdot\pi}{(1-Se)\cdot\pi+(1-\pi)}$
- ...and therefore

$$E[Y|m = 0] = \frac{(1 - Se) \cdot \pi \cdot E[Y|\mu = 1] + (1 - \pi) \cdot E[Y|\mu = 0]}{(1 - Se) \cdot \pi + (1 - \pi)}$$

- This is not equal to $E[Y|\mu = 0]$!

[9] Note that this is also equal to $1 - p(\mu = 0|m = 0)$, so $1 - NPV$.

# Consequences of ignoring the problem

- If you take the difference and simplify, you discover that the difference in means estimates:

$$\hat{\Delta} = (E[Y|\mu = 1] - E[Y|\mu = 0]) \frac{(1-\pi)}{(1-Se)\pi + (1-\pi)}$$

- This only equals the true quantity of interest $E[Y|\mu = 1] - E[Y|\mu = 0]$ if...
  - $Se = 1$ (no measurement error)
  - $\pi = 0$ (no cases)
- Otherwise, there will be attenuation bias towards 0 versus the truth, since the denominator $(1 - Se)\pi + (1 - \pi)$ increases as sensitivity decreases.

# Fixing the problem?

- The preceding calculation was only with one-sided measurement error ($Sp = 1, Se \neq 1$)
  - For the case where $Sp = 1, Se \neq 1$, just re-arrange the equation on the previous slide to get

$$E[Y|\mu = 1] - E[Y|\mu = 0] = \hat{\Delta}\frac{(1 - Se)\pi + (1 - \pi)}{(1 - \pi)}$$

- This gives a solution for bias correcting the difference in means in this specific case where we have one-sided measurement error
- It is more complicated (but not *better*) if there is error in both the $\mu = 0$ and $\mu = 1$ cases.
- However, this problem is solvable *only if* you know $Se$ and/or $Sp$ (depending on the type of measurement error) and $\pi$.

# A more general solution

- You could also simply run a regression for $Y$ where, instead of using the binary $m \in 0, 1$ as the explanatory variable, you use $p(m = 1)$.
  - In this instance, calculating $p(m = 1)$ does not require knowing $Se$ and/or $Sp$ and $\pi$.
- More generally:
  - If you have used something like a logistic regression model as the basis of your measurement strategy, you can just use the predicted probability $p(m = 1)$ directly from that measurement model!
- Bottom line:
  - If you can quantify the measurement uncertainty in a categorical variable, you should use the probability classification rather than the point prediction as your measure in subsequent analyses

# Summing up

- Sometimes, it makes sense to measure the concept of interest on a nominal or ordinal scale
  - This means that units are **classified** into categories instead of being placed on an interval scale
- **Supervised** class measurement can be done by:
  - Recoding scales into categories
  - Specifiying the indicators and aggregation such that it yields classes
  - Using training data with nominal/ordinal 'gold-standard' measurement
- Measurement error looks different when classifying stuff. We have:
  - Accuracy & Total error rate
  - Sensitivity & Specificity
  - Positive predictive value & Negative predictive value
- Whenever your measurement method yields predicted probabilities, it often makes sense to use those instead of the predicted class