Research design & study execution workshop series

Session 8

OCTOBER 6, 2015

Quick review of Sessions 1-7

- How to identify a “good” research question
- Common study designs: Pros & cons
- Selecting appropriate study subjects
- Understanding variables types and their measurement
- Good data management: Data collection, entry & cleaning

Case study: Football-related injuries

Nuts and bolts of good data management: Part III

Data recoding and archiving

Data management process

All of the steps required to create a clean data set ready to be analyzed

Overview of the process

1. Collect the data
2. Enter the data
3. Clean the data
4. Create and format new variables
5. Document and archive all data sets

Create and format new variables in order to answer your research question(s)

Data management step	**End product**
1. Data collection and entry	Raw data set
2. Data cleaning	Clean data set
3. Create all new variables required for analysis	Analytical data set

Why generate new variables?

1. Create meaningful groups (cutoffs)
2. Change codes to make an analysis possible
3. Reverse the order of a multipoint scale
4. Combine groups to avoid sparse data
5. Create variables that reflect change

Best practices

- Never delete the raw variables (always add new ones)
- Use meaningful names
- Always cross-check to ensure recoding process worked
- Document what you did, how you did it & why
- Consider consulting a data analyst for projects that require complicated data set manipulation

Five worked examples

1. Create meaningful groups (use cutoffs)

Problem

You have a continuous variable

Situation 1. You need to identify normal vs. abnormal results

Situation 2. You need to identify 3 groups (low, medium, high)

Situation 1: Identify normal vs. abnormal

You recorded the maximal outer diameter (MOD) of the appendix as a continuous variable with values from 1-20 mm

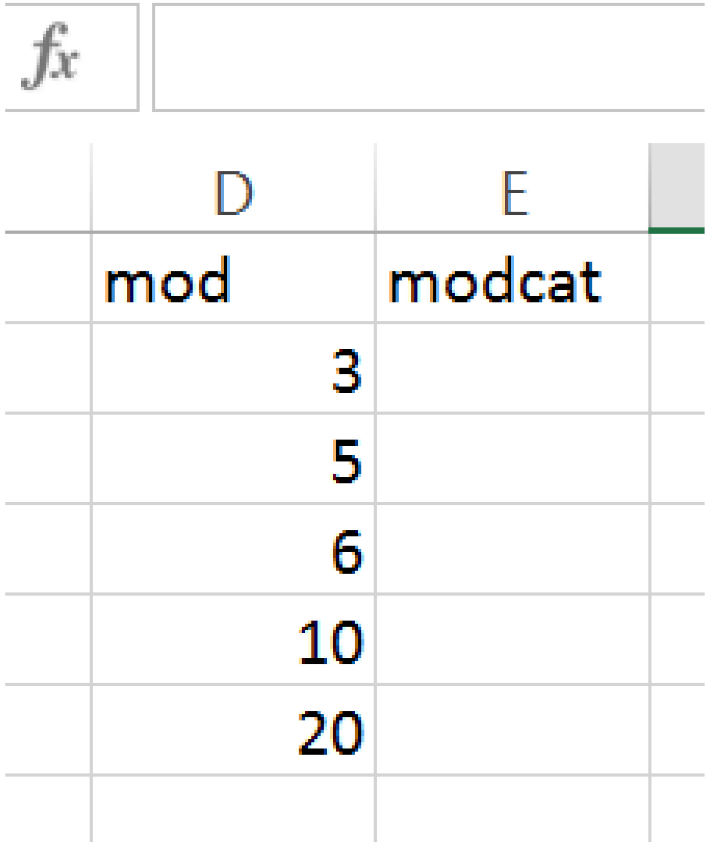
You need to categorize patients as:

1-6 mm = “normal”

7-20 mm = “abnormal”

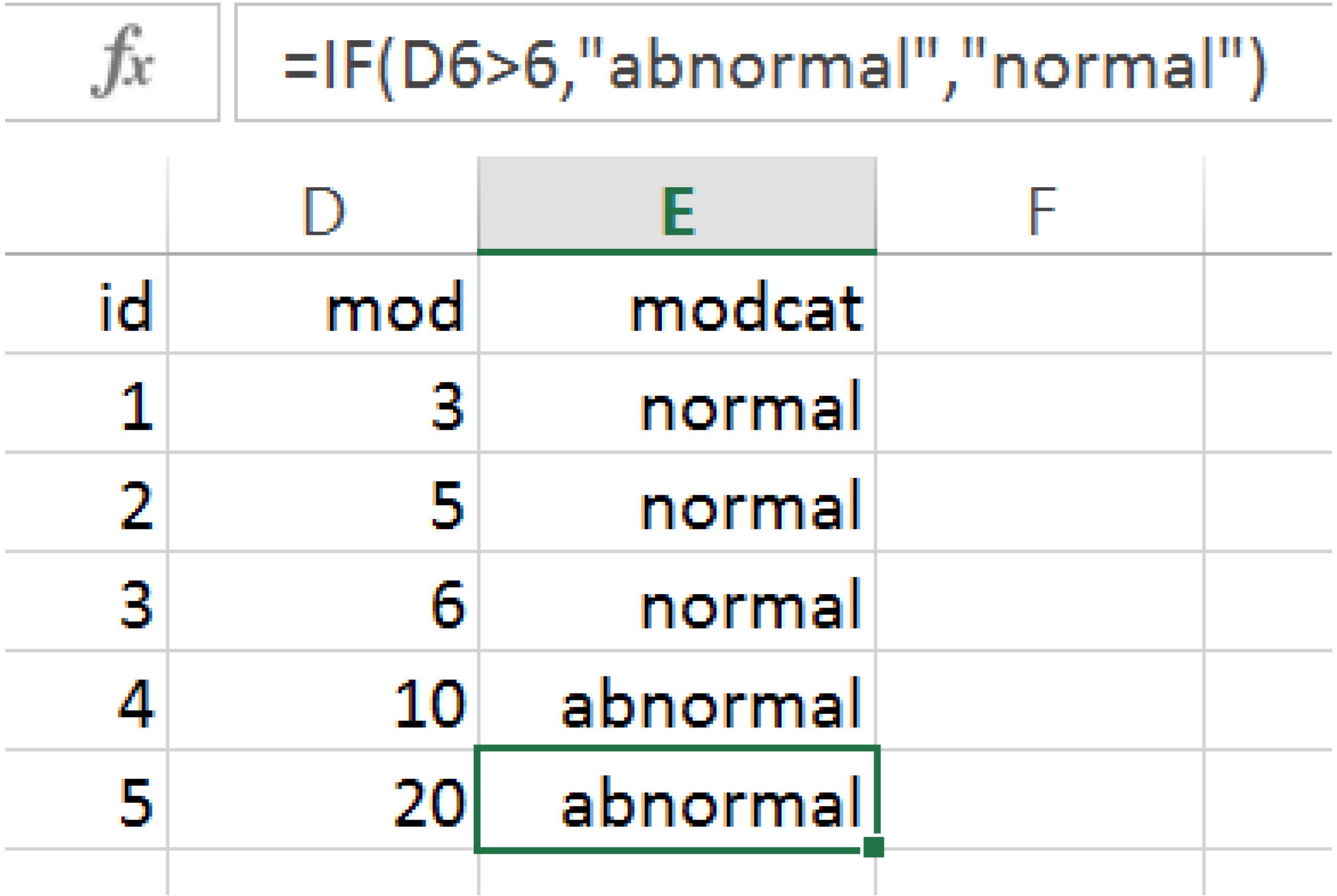
Use the “IF” function in Excel

1. Add a blank column next to “mod” and name it “modcat”



D	E
mod	modcat
3	
5	
6	
10	
20	

2. Use the “IF” function to create 2 mutually exclusive groups



```
=IF(D6>6,"abnormal","normal")
```

	D	E	F
id	mod	modcat	
1	3	normal	
2	5	normal	
3	6	normal	
4	10	abnormal	
5	20	abnormal	

Situation 2: Identify three groups

You have patient pain scores recorded a continuous variable from 1-10

You need to categorize pain scores as:

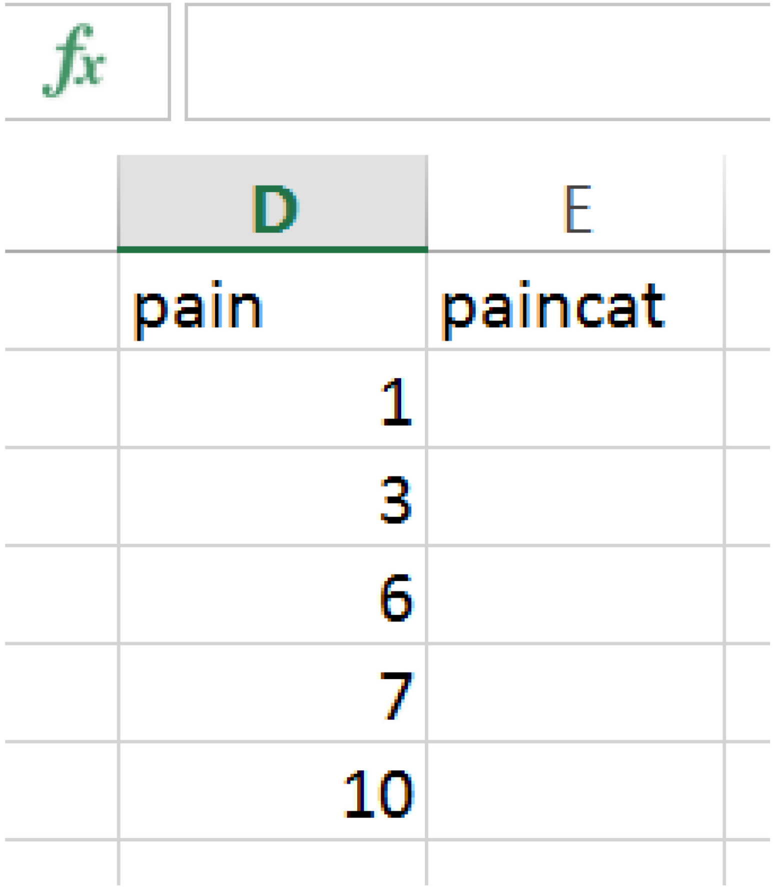
1-3 = “low”

4-7 = “medium”

8-10 = “high”

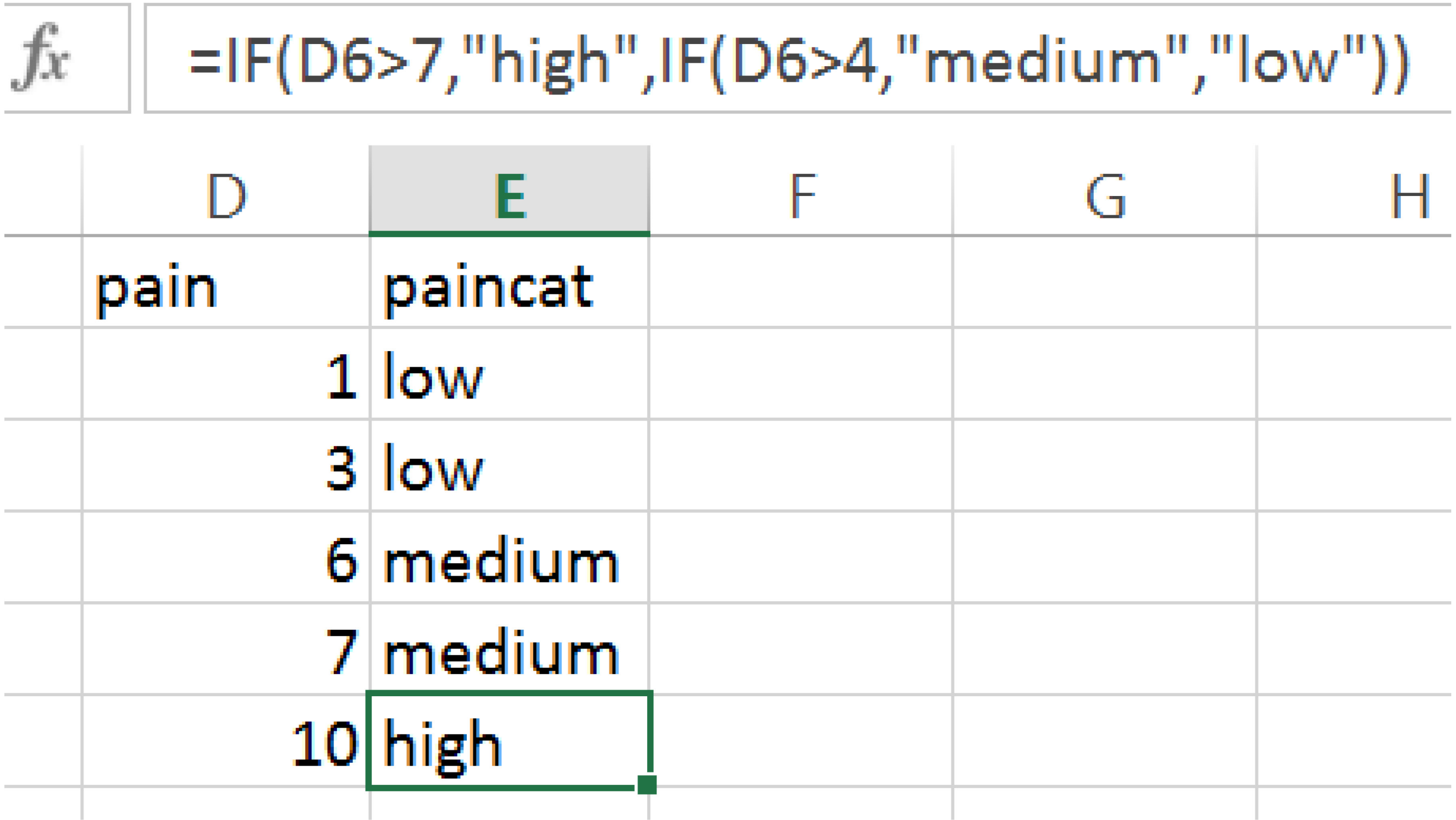
Use nested “IF” functions

1. Add a blank column next to “pain” and name it “paincat”



D	E
pain	paincat
1	
3	
6	
7	
10	

2. Use the “IF” function to create 3 mutually exclusive groups



```
=IF(D6>7,"high",IF(D6>4,"medium","low"))
```

D	E	F	G	H
pain	paincat			
1	low			
3	low			
6	medium			
7	medium			
10	high			

2. Change codes to make an analysis possible

Problem

Some statistical analyses require a dichotomous variable coded as 0=no/absent; 1=yes/present

Situation 1. You have a 1=yes; 2=no

Situation 2. You have a text variable indicating “yes” or “no”

Situation 3. You have other information that leads to yes/no

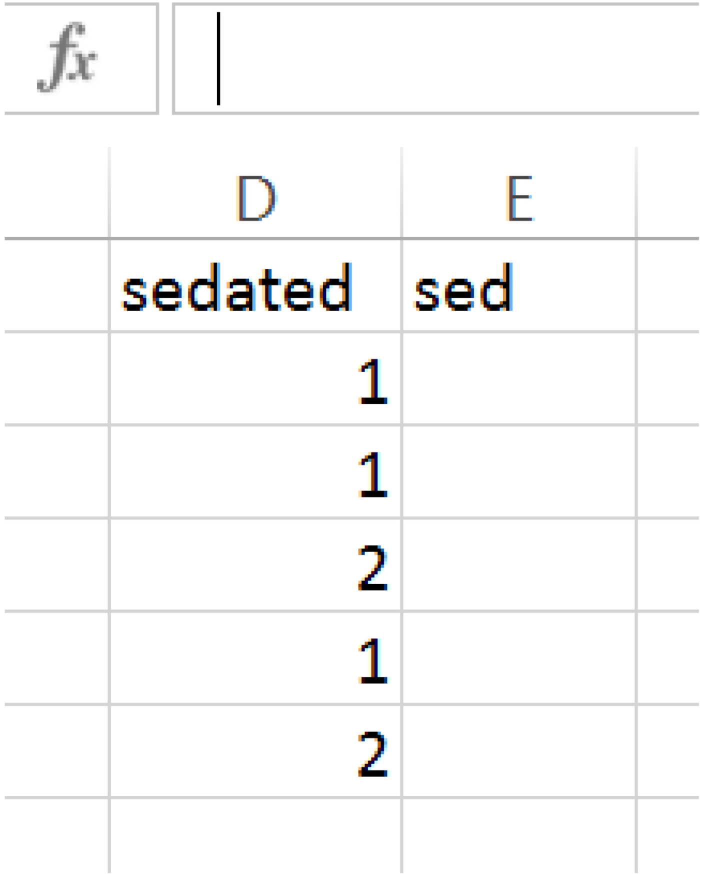
Situation 1: Reverse numeric coding

You recorded patient sedation status as 1=yes; 2=no

You need 1=yes; 0=no

Use the “IF” function

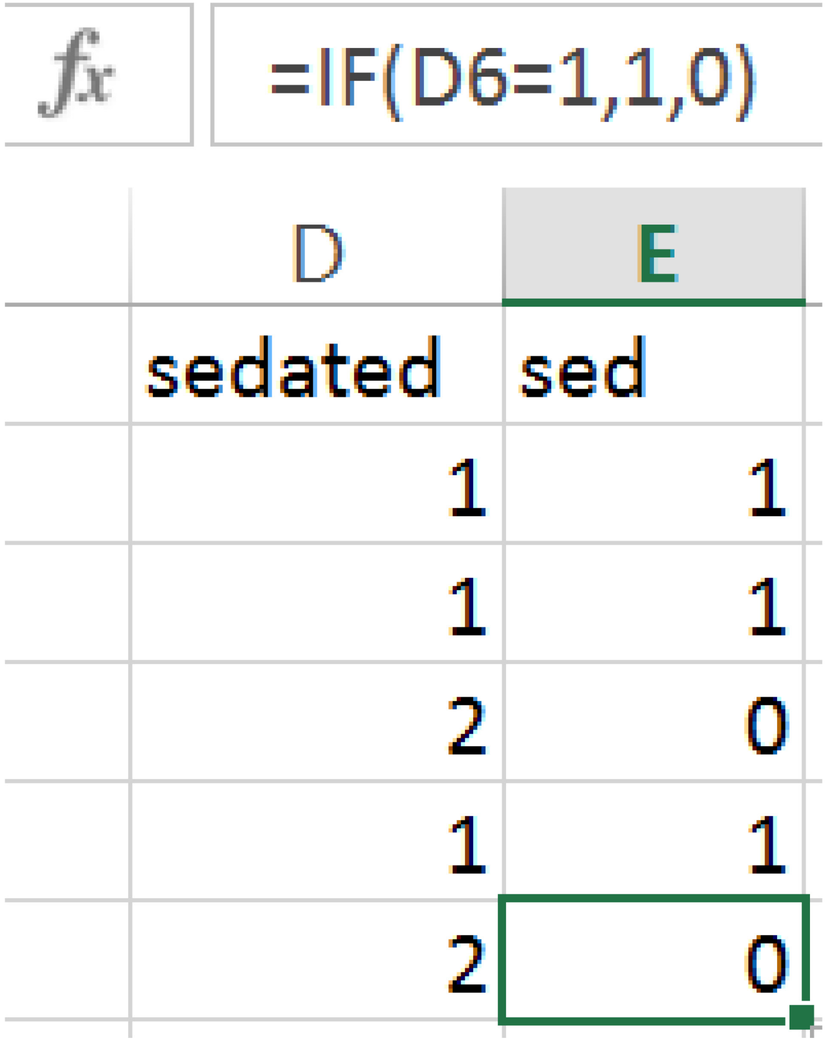
1. Add a blank column next to “sedated” and name it “sed”



D	E
sedated	sed
1	
1	
2	
1	
2	

2. Use the “IF” function to create sed = 1 (for yes)

sed = 0 (for no)



=IF(D6=1,1,0)

D	E
sedated	sed
1	1
1	1
2	0
1	1
2	0

Situation 2: Text-to-numeric conversions

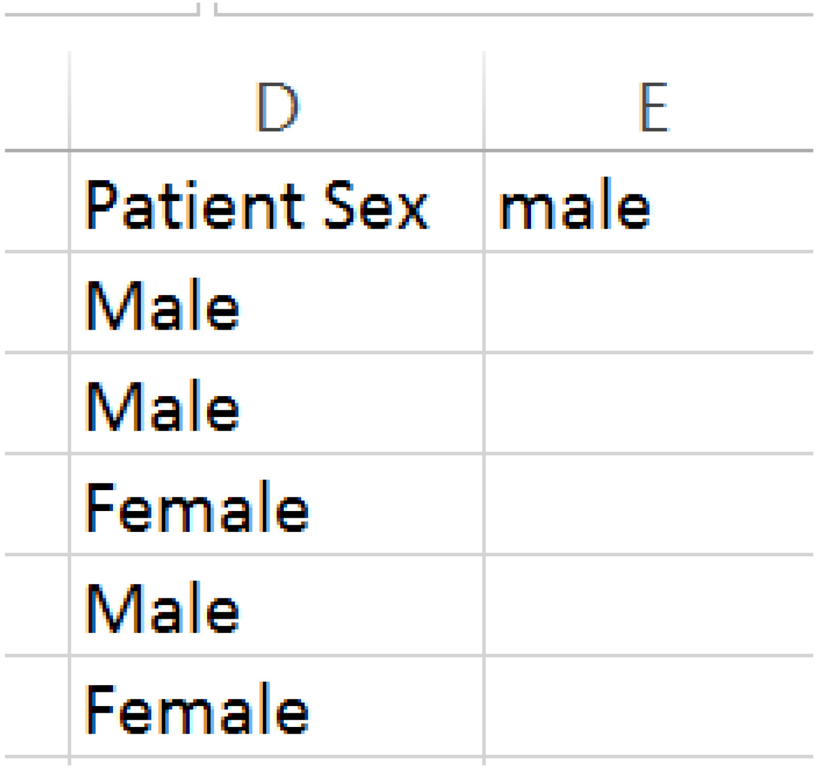
The question to answer is “Is the patient male?”

You have patient sex recorded as Male or Female

You need Female=0; Male=1

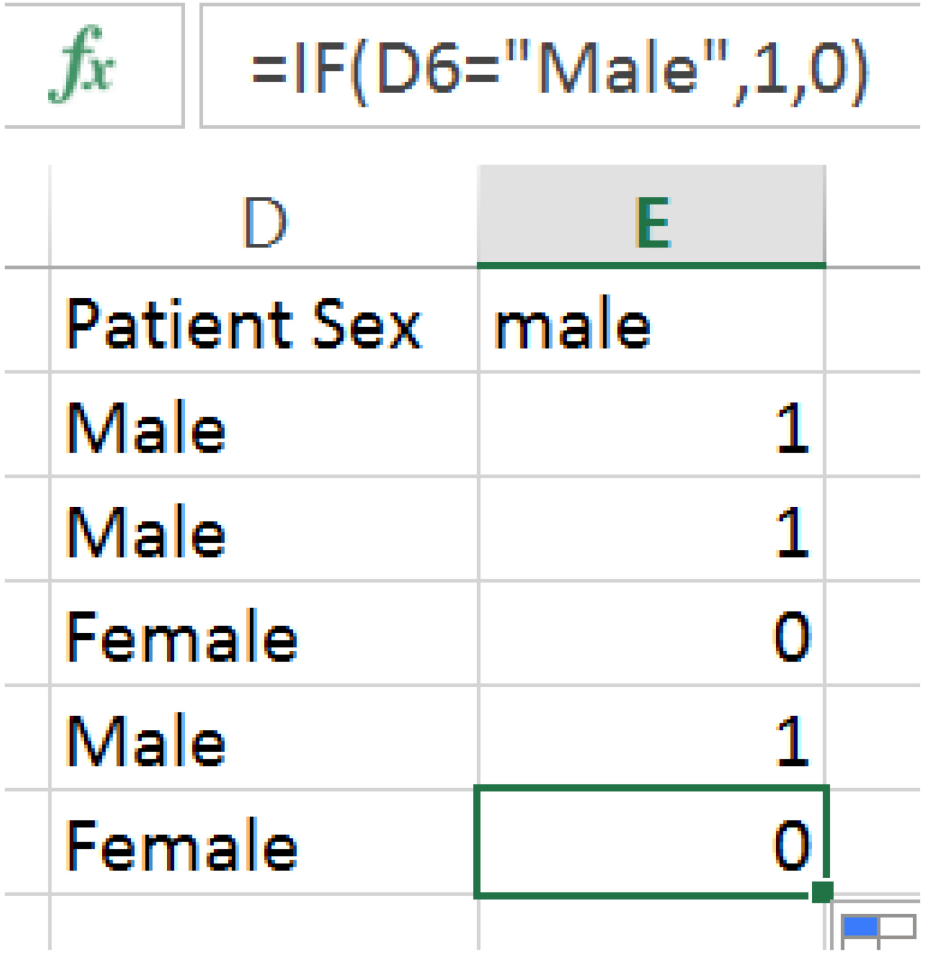
Use the “IF” function

1. Add a blank column next to “Patient Sex” and name it “male”



D	E
Patient Sex	male
Male	
Male	
Female	
Male	
Female	

2. Use the “IF” function to create male = 1 and female = 0



```
=IF(D6="Male",1,0)
```

D	E
Patient Sex	male
Male	1
Male	1
Female	0
Male	1
Female	0

Situation 3: Creating yes/no categories

NPO example

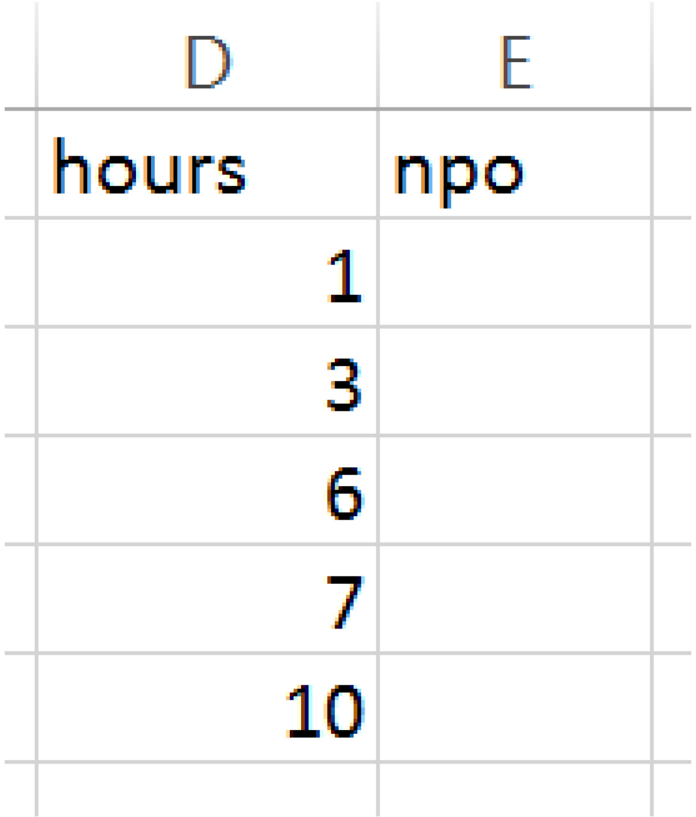
Patient is schedule for a sedated MRI exam and needs to be NPO for 6 (or more) hours to be sedated

You ask:

How many hours ago did he or she last eat or drink?

Use the “IF” function

1. Add a blank column next to “hours” and name it “npo”

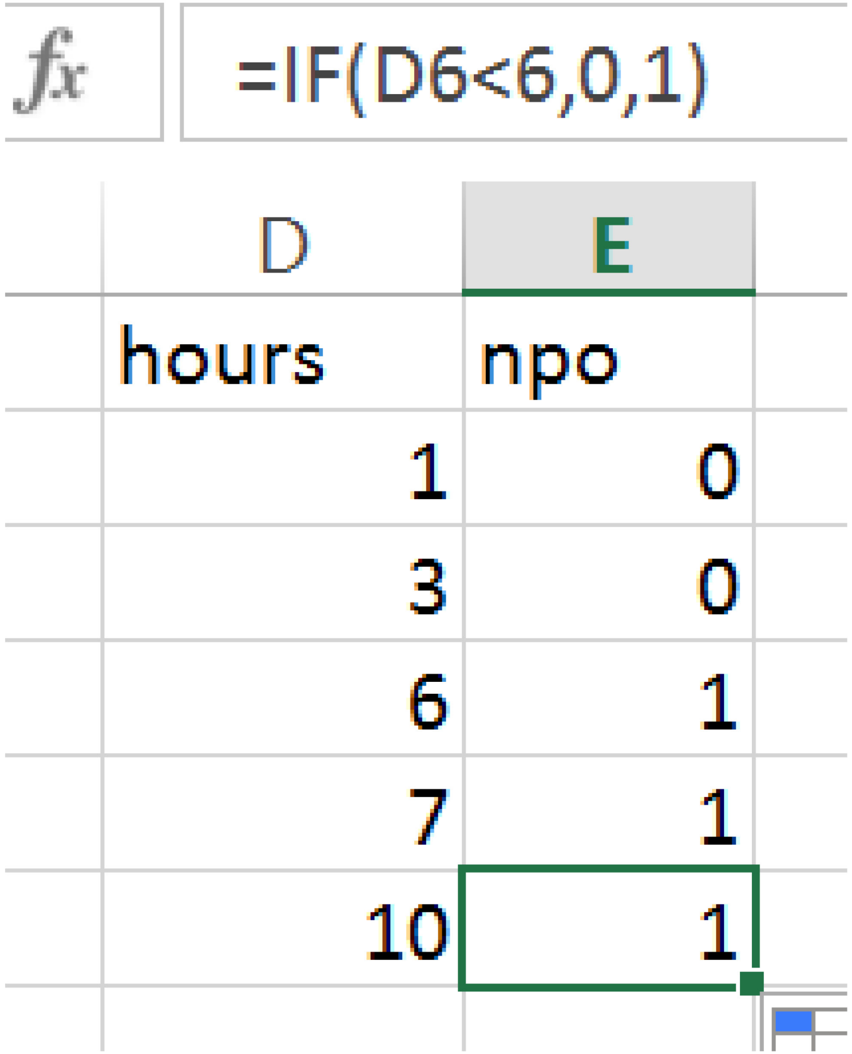


D	E
hours	npo
1	
3	
6	
7	
10	

2. Use the “IF” function to test the value of hours

Make npo=0 if hours <6

Make npo=1 if hours >=6



```
=IF(D6<6,0,1)
```

D	E
hours	npo
1	0
3	0
6	1
7	1
10	1

3. Reverse the order of a multipoint scale

Problem

You have a 5-point scale

1 “strongly agree”

2 “agree”

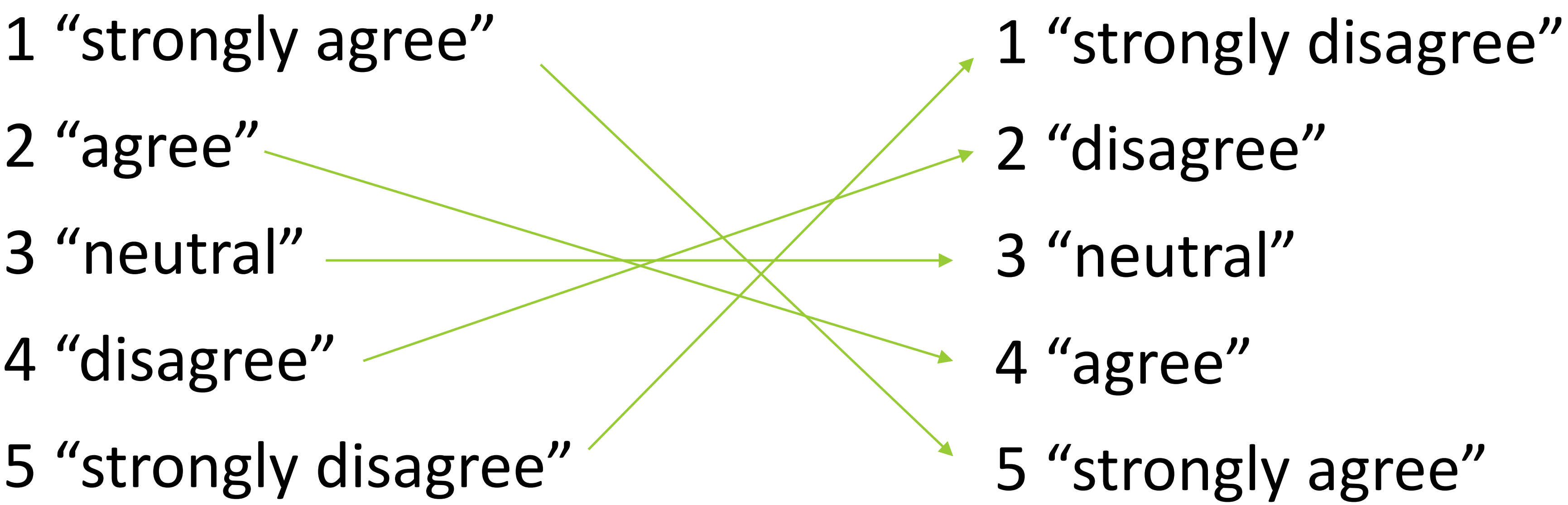
3 “neutral”

4 “disagree”

5 “strongly disagree”

3. Reverse the order of a multipoint scale

Problem



You have a 5-point scale

1 “strongly agree”

2 “agree”

3 “neutral”

4 “disagree”

5 “strongly disagree”

You need…

1 “strongly disagree”

2 “disagree”

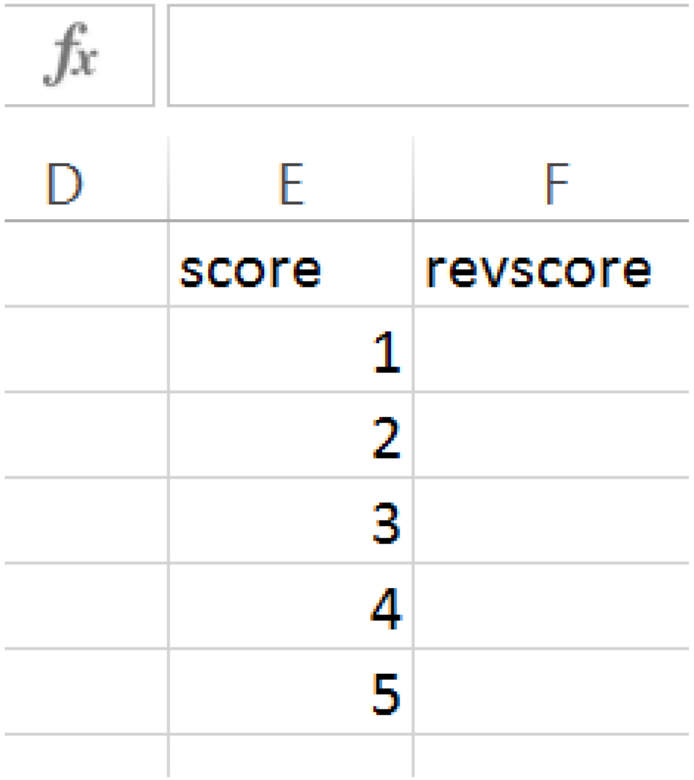
3 “neutral”

4 “agree”

5 “strongly agree”

Use addition and subtraction in Excel

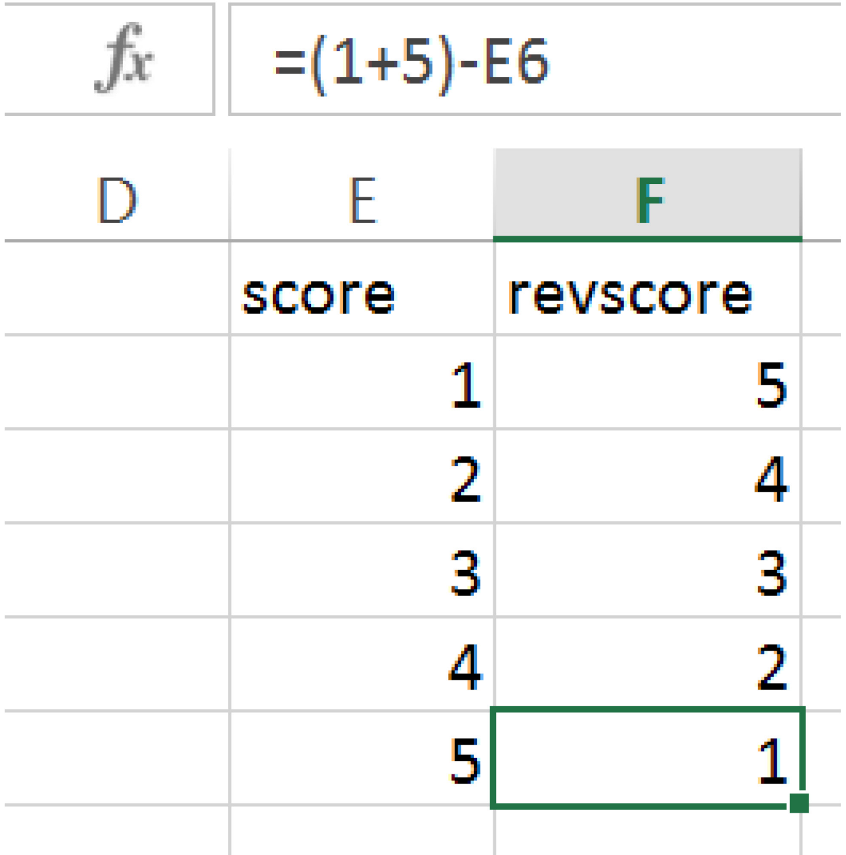
1. Add a blank column next to “score” and name it “revscore”
2. Determine the minimum and maximum values of your revised score [here 1 & 5]



D	E	F
	score	revscore
	1	
	2	
	3	
	4	
	5	

3. Use this equation to create "revscore"

Reversed score = (minimum + maximum) - original value



fx =(1+5)-E6

D	E	F
	score	revscore
	1	5
	2	4
	3	3
	4	2
	5	1

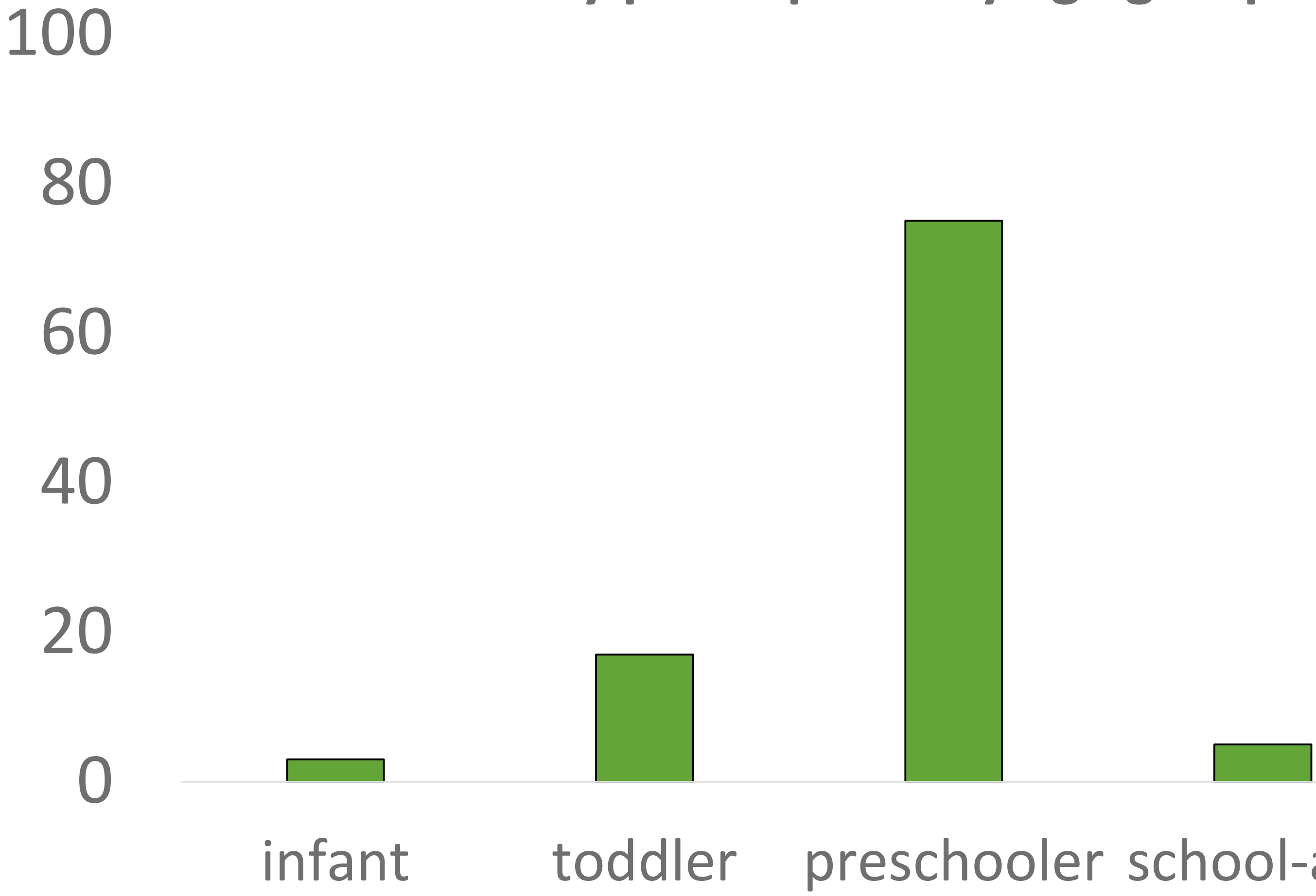
4. Combine groups to avoid sparse data

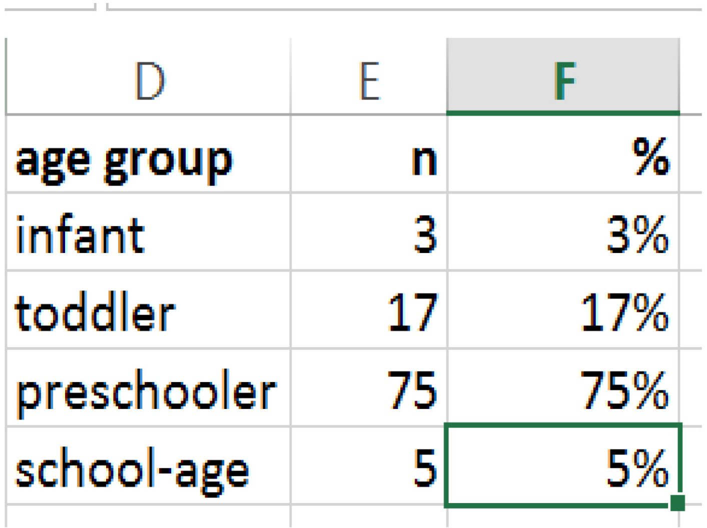
Problem

Certain statistical tests require a minimum sample size

You have far fewer observations than expected in certain categories

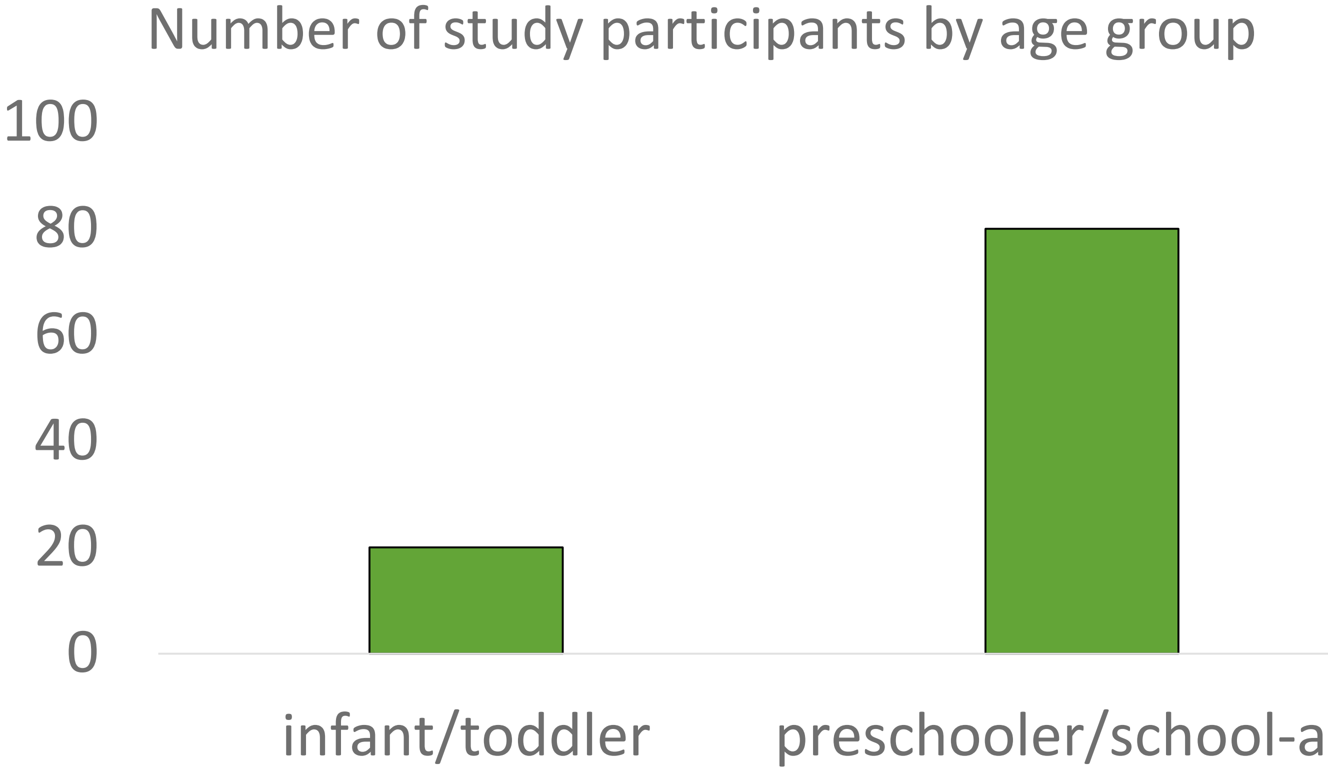
Number of study participants by age group

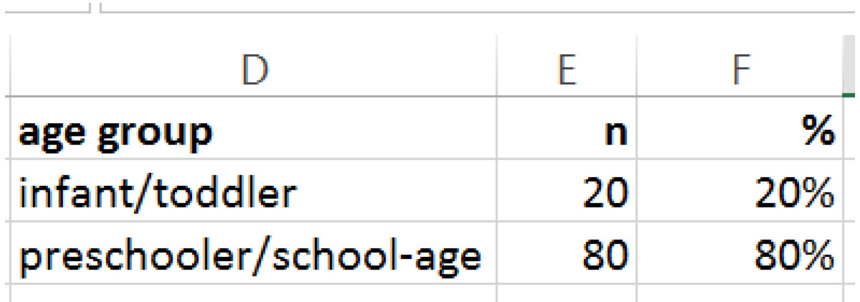




age group	n	%
infant	3	3%
toddler	17	17%
preschooler	75	75%
school-age	5	5%

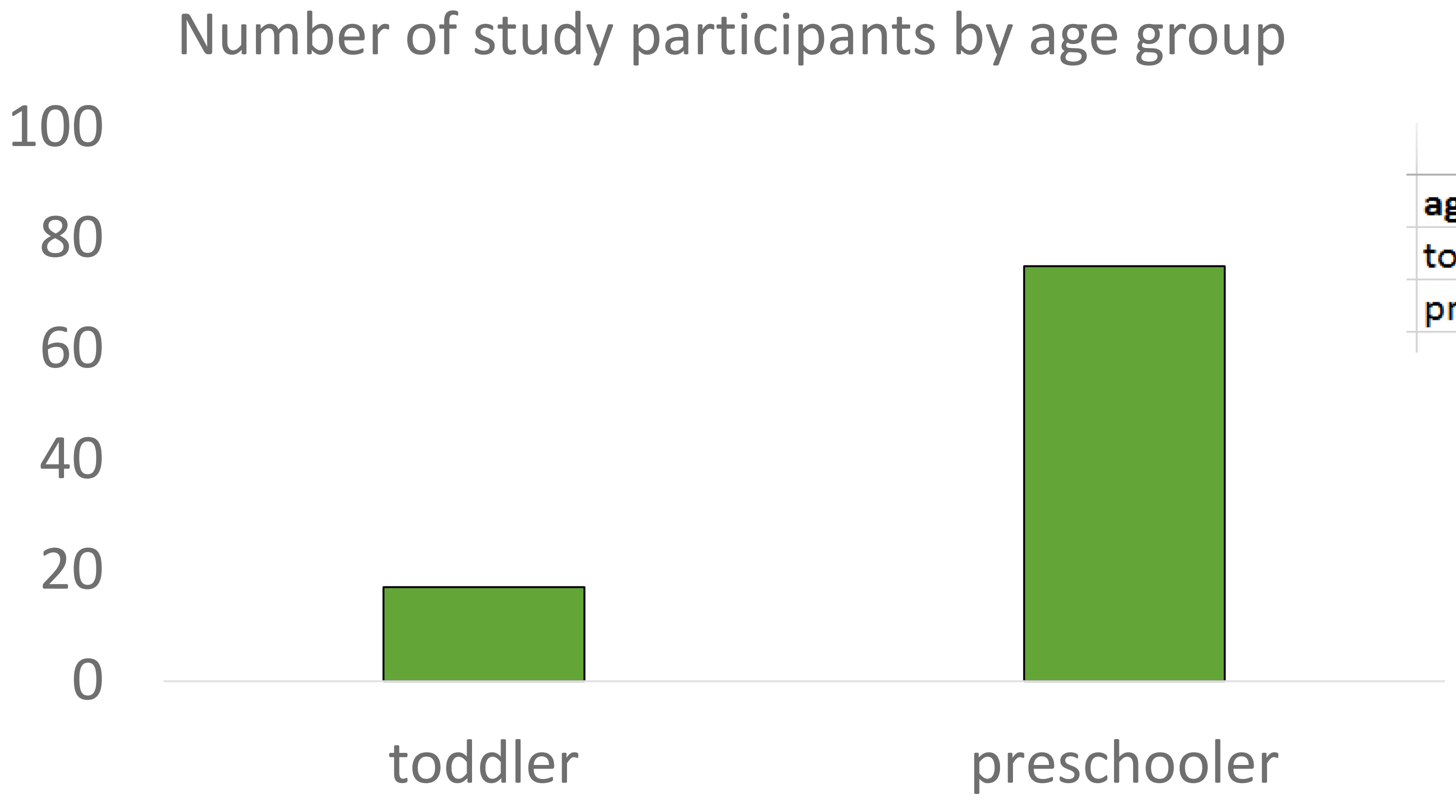
Option 1. Combine categories

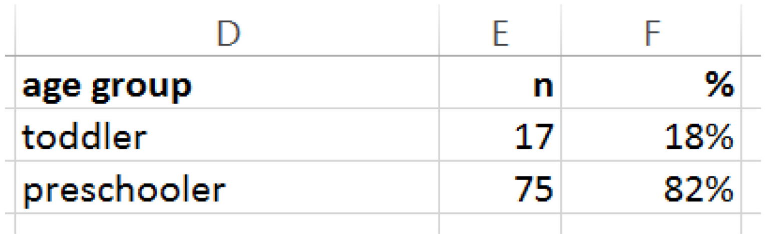




D	E	F
age group	**n**	**%**
infant/toddler	20	20%
preschooler/school-age	80	80%

Option 2. Drop categories





D	E	F
age group	**n**	**%**
toddler	17	18%
preschooler	75	82%

5. Create variables that reflect change

Problem

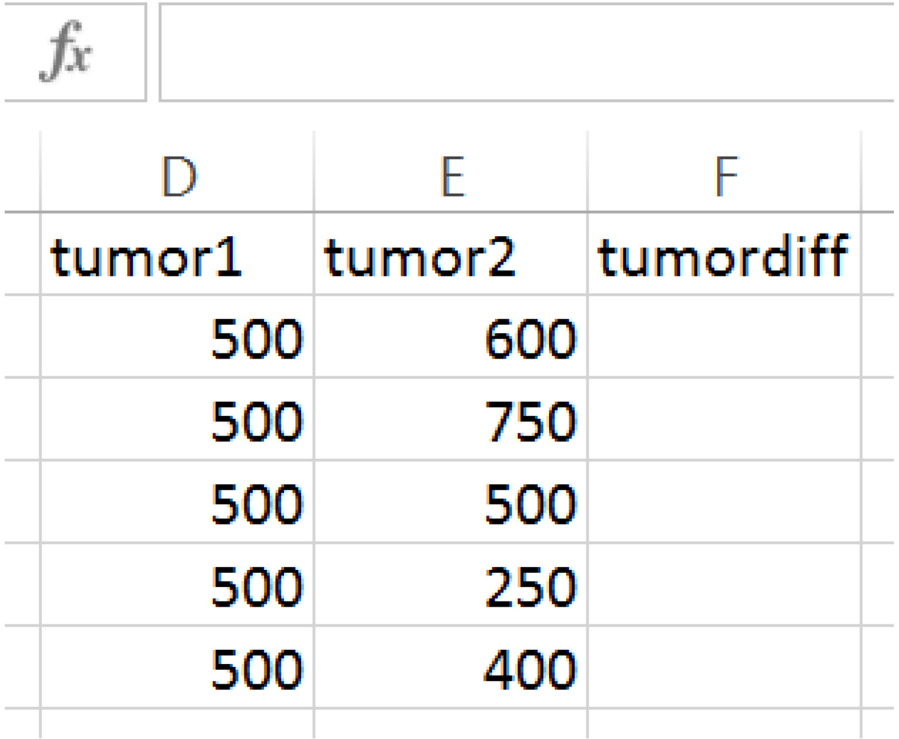
You have beginning and ending values, but you need to measure change over time

Situation 1. Size of anatomical feature before and after treatment = response to treatment

Situation 2. Date of birth and date of exam = Age at exam

Situation 1. Use subtraction in Excel

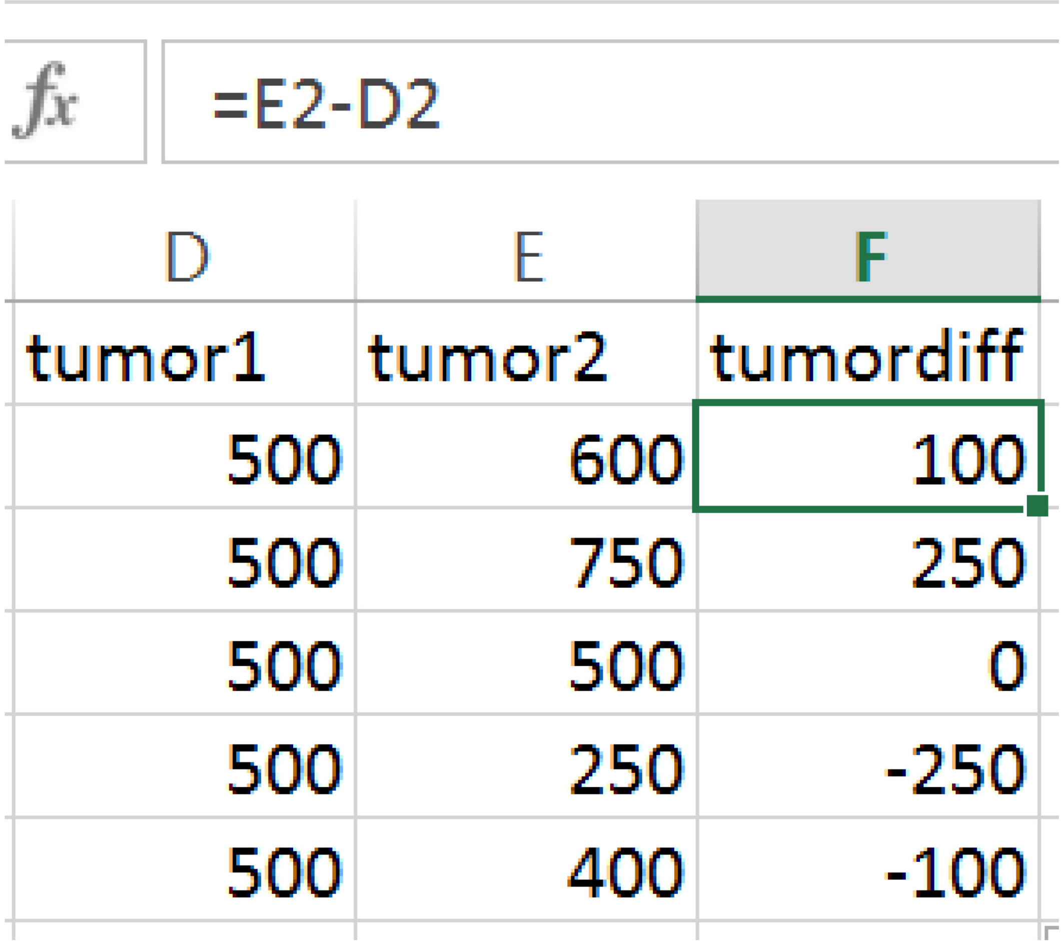
1. Add a blank column next to "tumor2" and name it "tumordiff"



tumor1	tumor2	tumordiff
500	600	
500	750	
500	500	
500	250	
500	400	

3. Use subtraction to calculate absolute change

Change in size of tumor = (tumor2 – tumor1)

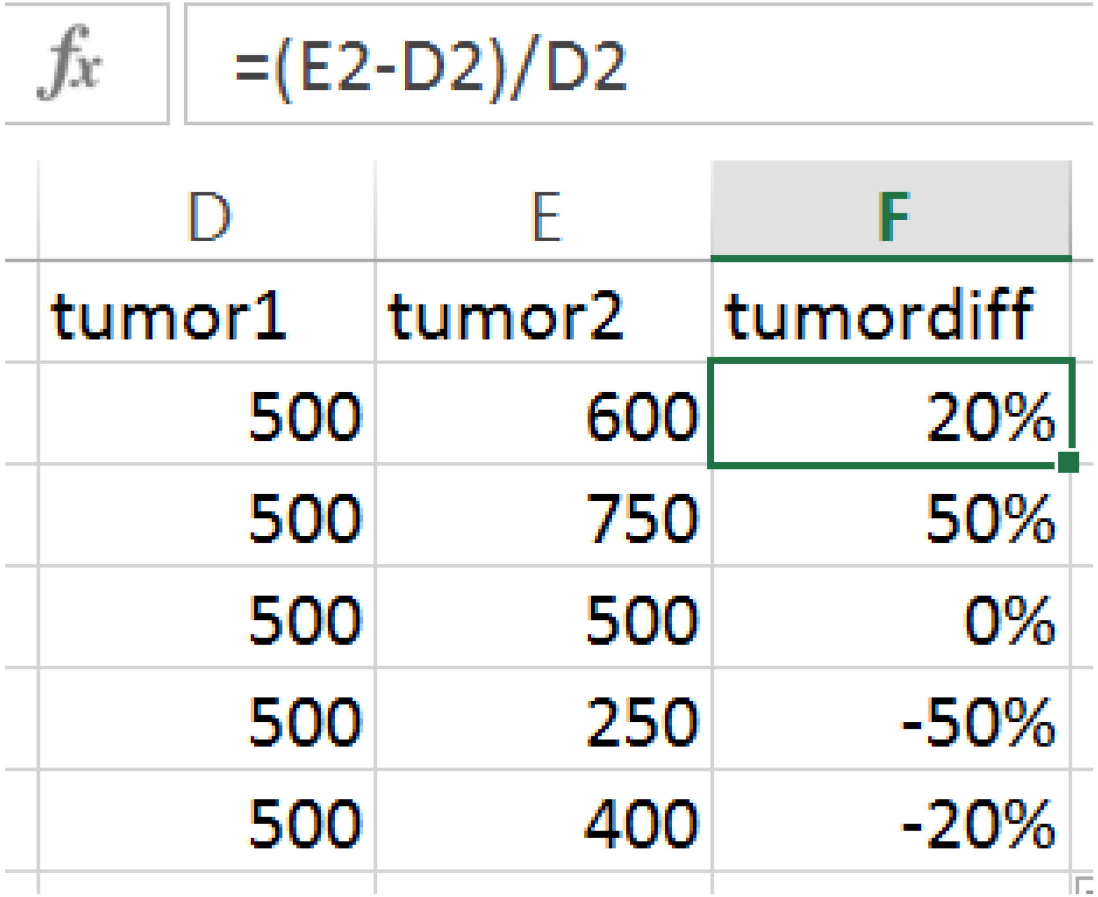


=E2-D2

D	E	F
tumor1	tumor2	tumordiff
500	600	100
500	750	250
500	500	0
500	250	-250
500	400	-100

4. OR subtraction and multiplication (and format tumordiff as Percentage) to calculate relative change

% change in size of tumor = (tumor2 – tumor1)/tumor1

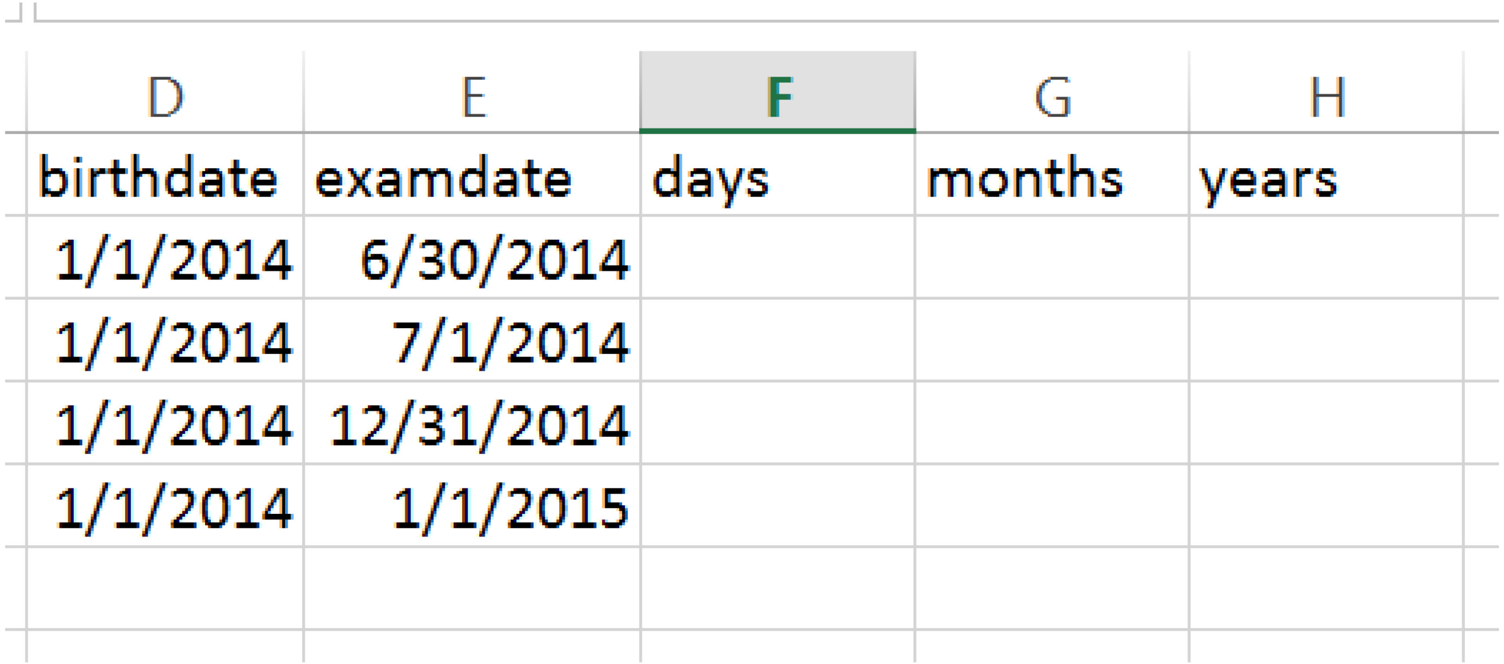


=(E2-D2)/D2

tumor1	tumor2	tumordiff
500	600	20%
500	750	50%
500	500	0%
500	250	-50%
500	400	-20%

Situation 2. Use “DATEDIF” function in Excel

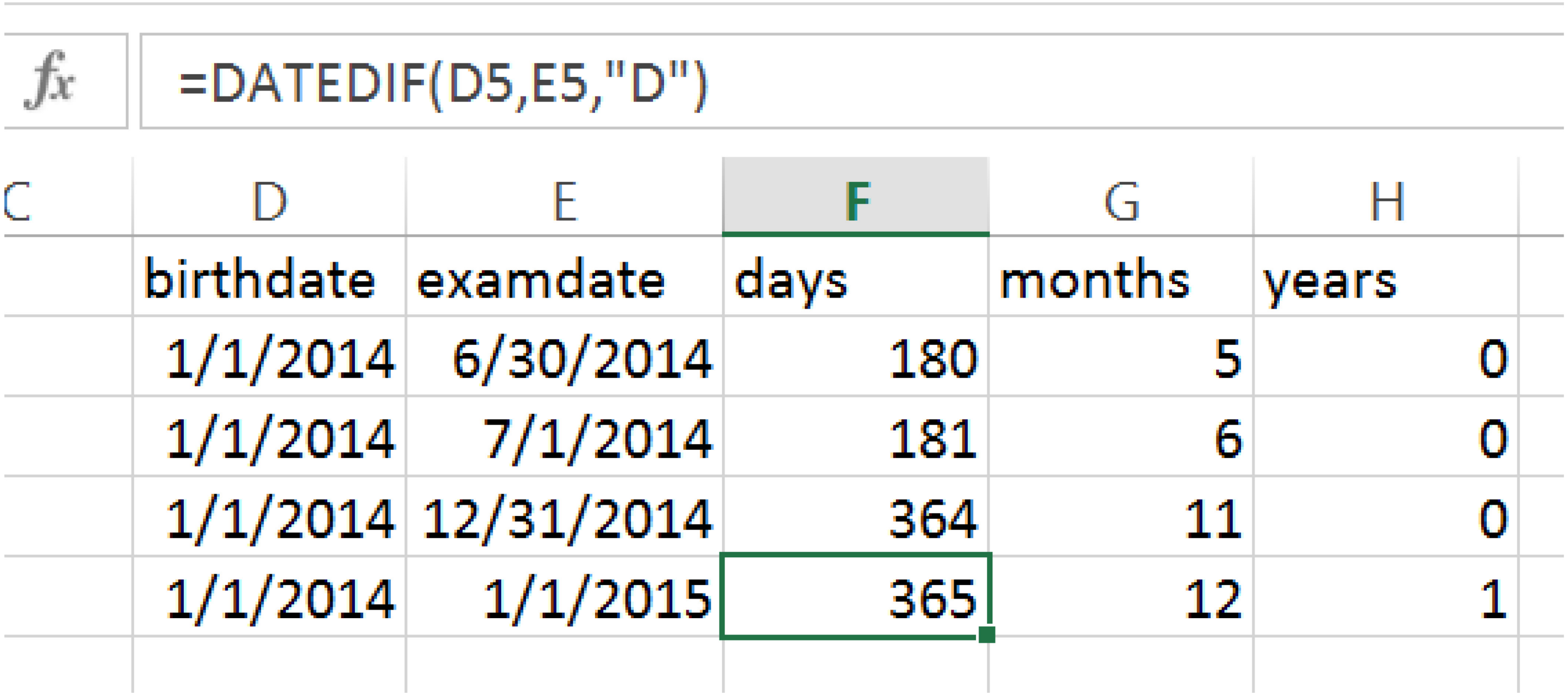
1. Add blank columns next to “examdate” and name them “days”, “months”, “years”



D	E	F	G	H
birthdate	examdate	days	months	years
1/1/2014	6/30/2014			
1/1/2014	7/1/2014			
1/1/2014	12/31/2014			
1/1/2014	1/1/2015			

2. To calculate age (in DAYS) use

=DATEDIF(D5,E5,”D”)

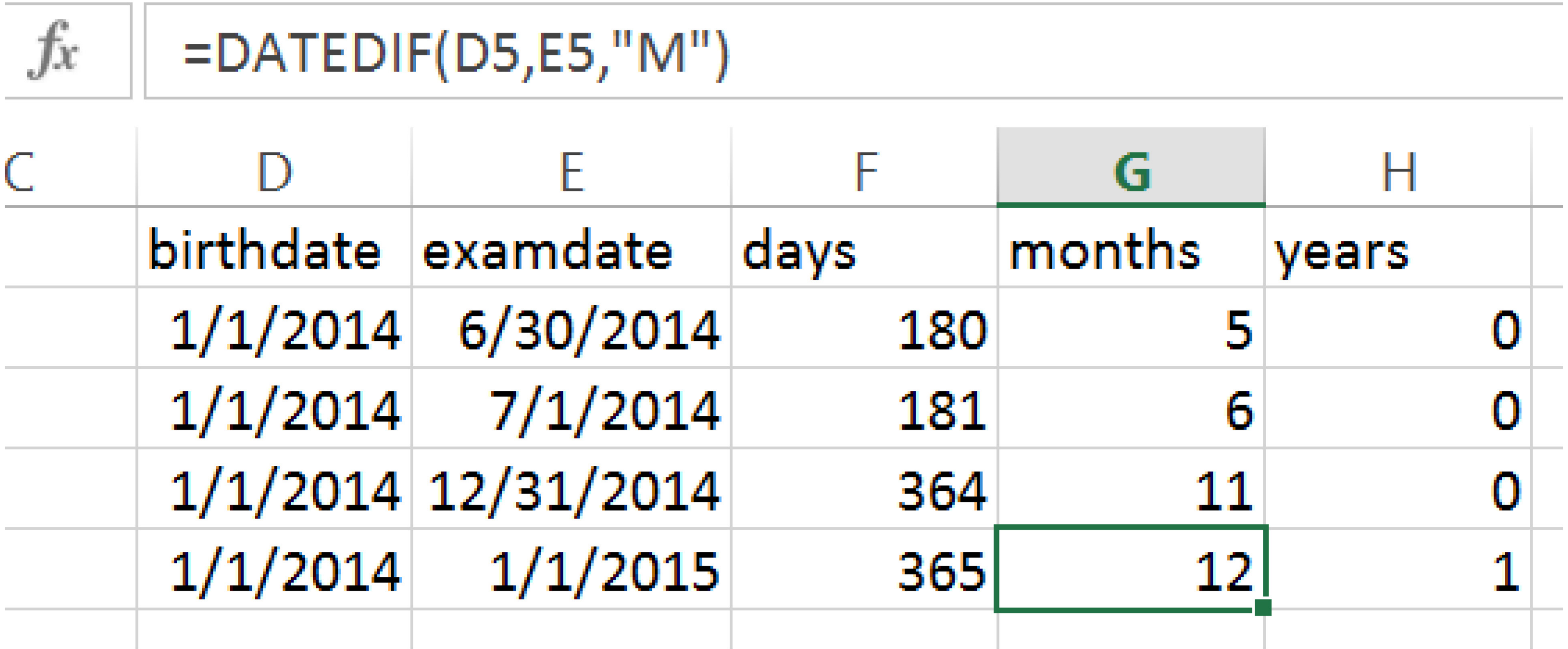


```
=DATEDIF(D5,E5,"D")
```

C	D	E	F	G	H
	birthdate	examdate	days	months	years
	1/1/2014	6/30/2014	180	5	0
	1/1/2014	7/1/2014	181	6	0
	1/1/2014	12/31/2014	364	11	0
	1/1/2014	1/1/2015	365	12	1

2. To calculate age (in total number of MONTHS elapsed) use

=DATEDIF(D5,E5,”M”)

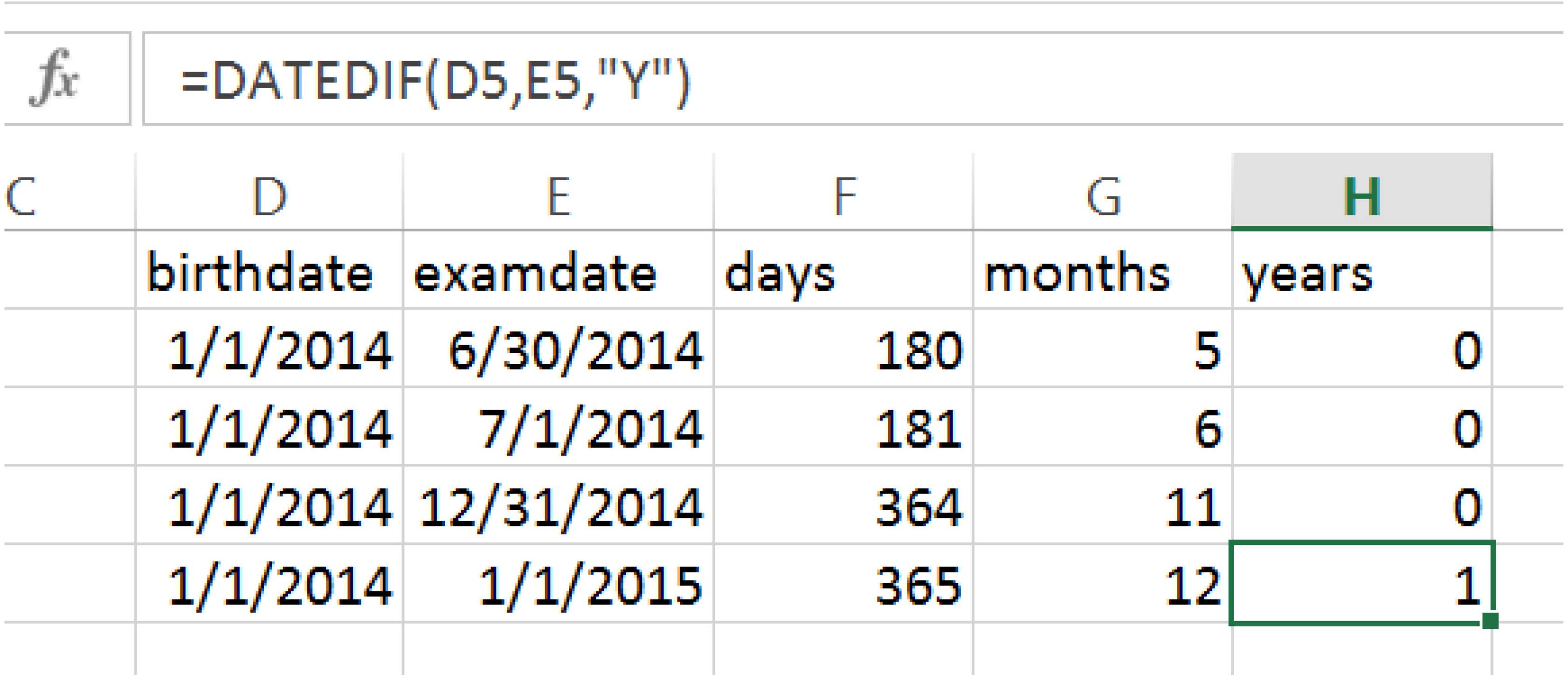


=DATEDIF(D5,E5,"M")

C	D	E	F	G	H
	birthdate	examdate	days	months	years
	1/1/2014	6/30/2014	180	5	0
	1/1/2014	7/1/2014	181	6	0
	1/1/2014	12/31/2014	364	11	0
	1/1/2014	1/1/2015	365	12	1

2. To calculate age (in total number of YEARS elapsed) use

=DATEDIF(D5,E5,”Y”)



fx =DATEDIF(D5,E5,"Y")

C	D	E	F	G	H
	birthdate	examdate	days	months	years
	1/1/2014	6/30/2014	180	5	0
	1/1/2014	7/1/2014	181	6	0
	1/1/2014	12/31/2014	364	11	0
	1/1/2014	1/1/2015	365	12	1

Common data formats

- Number
- Date
- Text
- Time

For example…

1. Set all missing data codes to ‘missing’
2. Format date variables as dates, numeric variables as numeric, etc.
3. Label all variables and categorical values so you don’t have to keep looking them up

Reasons to format variables properly

1. So your software works with them correctly
2. To save time during analysis & interpretation

Data archiving

Save backup copies of all key data files and important notes in order to protect your work

Best practices

1. Use systematic & reproducible methods
2. Archive all key files
   - Raw data, clean data, analytical data files
   - All data cleaning and recoding notes
3. Consider working with a data analyst on projects that require complex data manipulation

Questions or comments?

Next week

Basic data visualization techniques