Everything You Never Thought You Wanted to Know About Thought You Wanted to Know About Biostatistics & Study Design Services


Janet Foote, PhD, Assistant Professor, Public Health


Chengcheng Hu, PhD, Associate Professor, Public Health

Biostatistician

- Chengcheng Hu

Epidemiologist

- Janet Foote

THE UNIVERSITY OF ARIZONA | MEL AND ENID ZUCKERMAN COLLEGE OF PUBLIC HEALTH

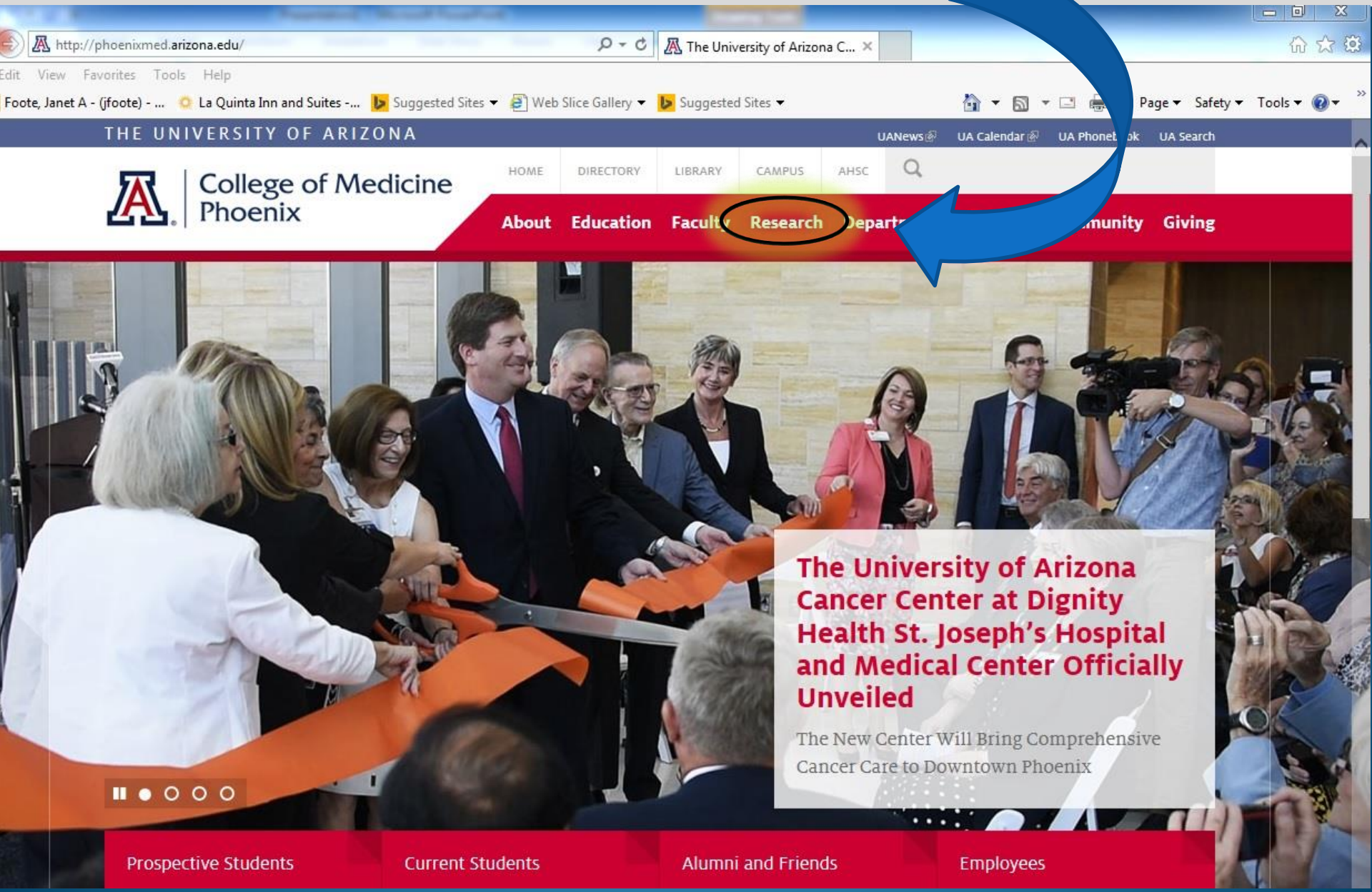
M.S.: Paul Kang

GRA: David Margraf / Shafquat “Shaf” Saif

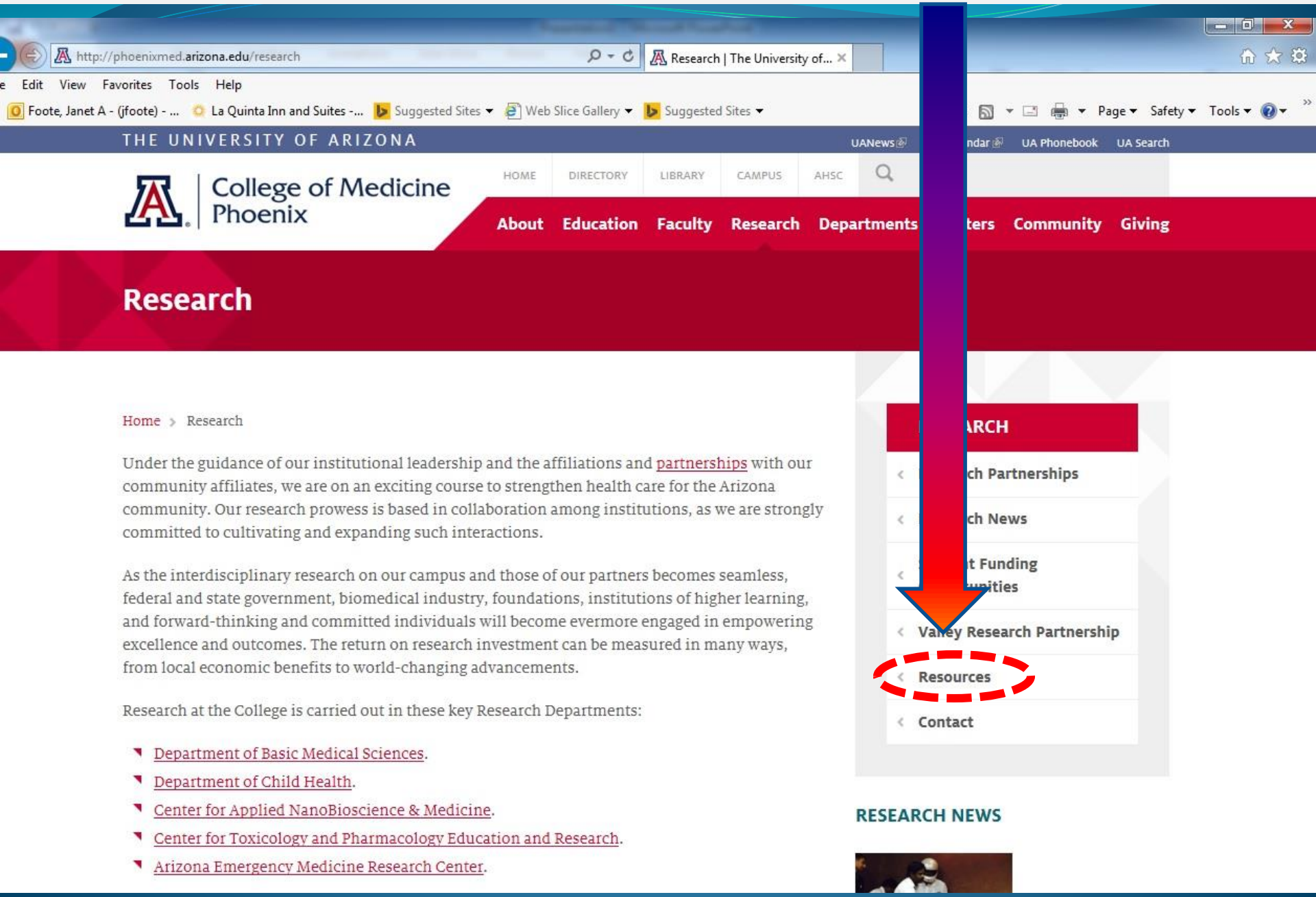
Additional Faculty from the Department of Biostatistics & Epidemiology as needed.

Finding more information online

Research

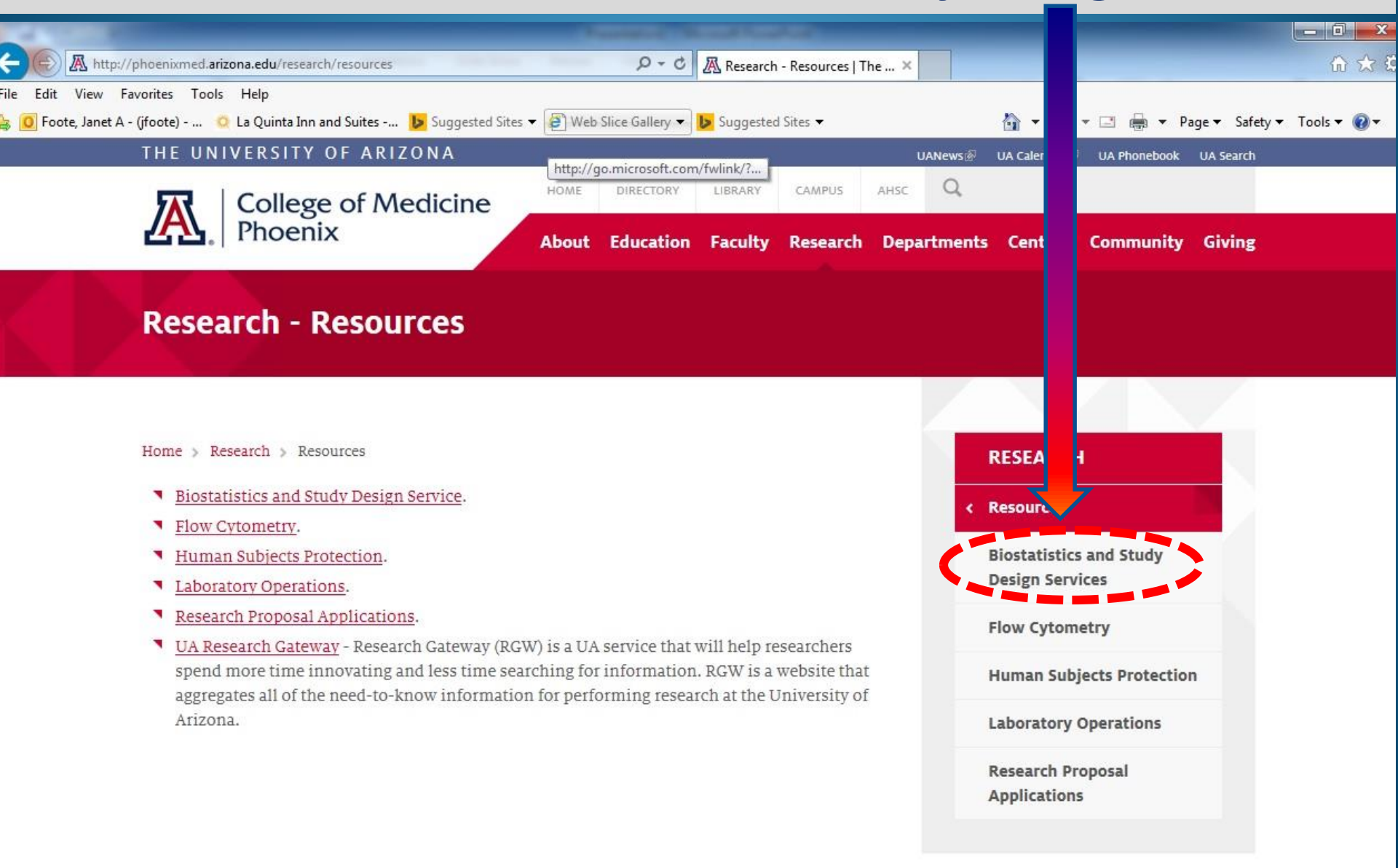


On the Research page, select **Resources**



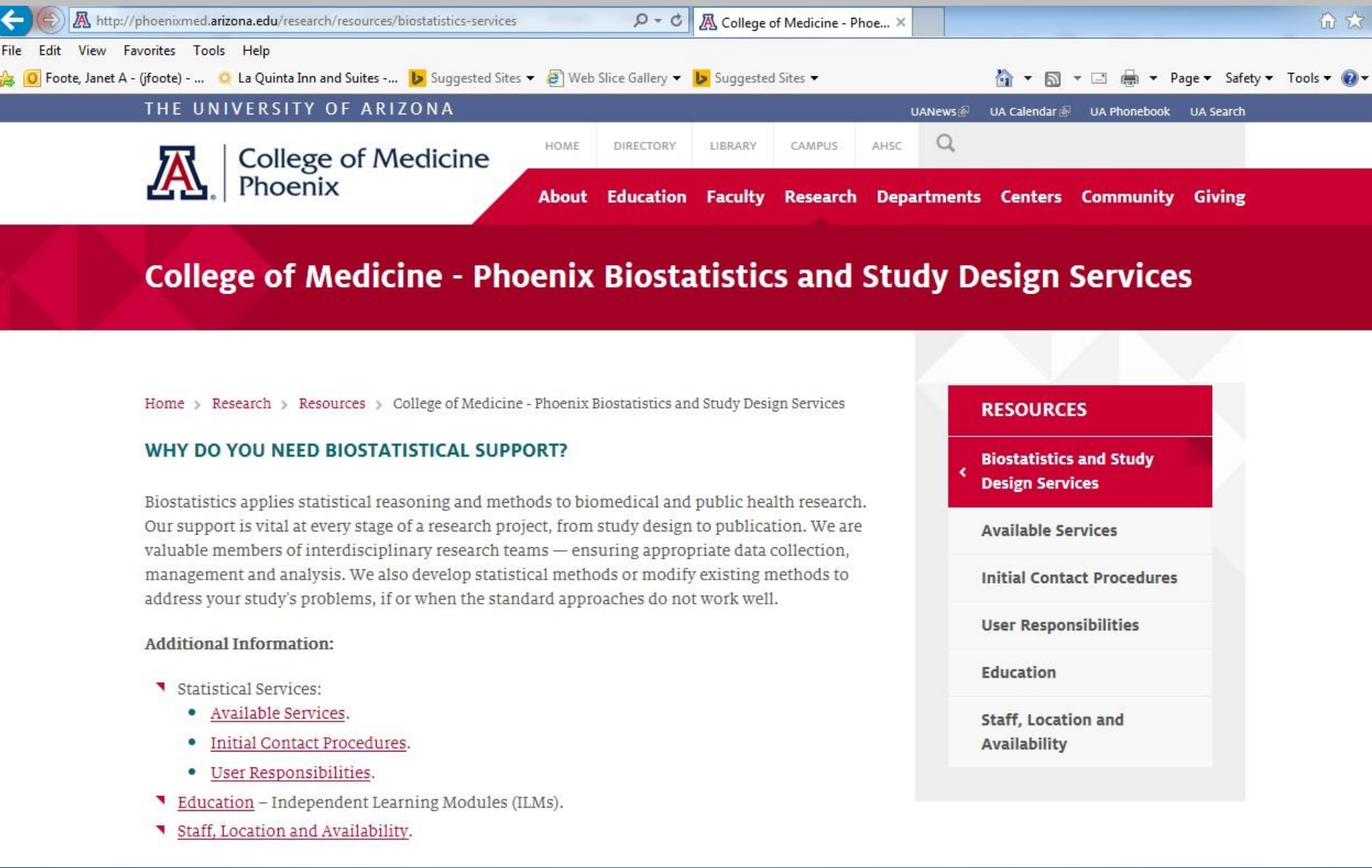
Under Resources, select

Biostatistics & Study Design Services

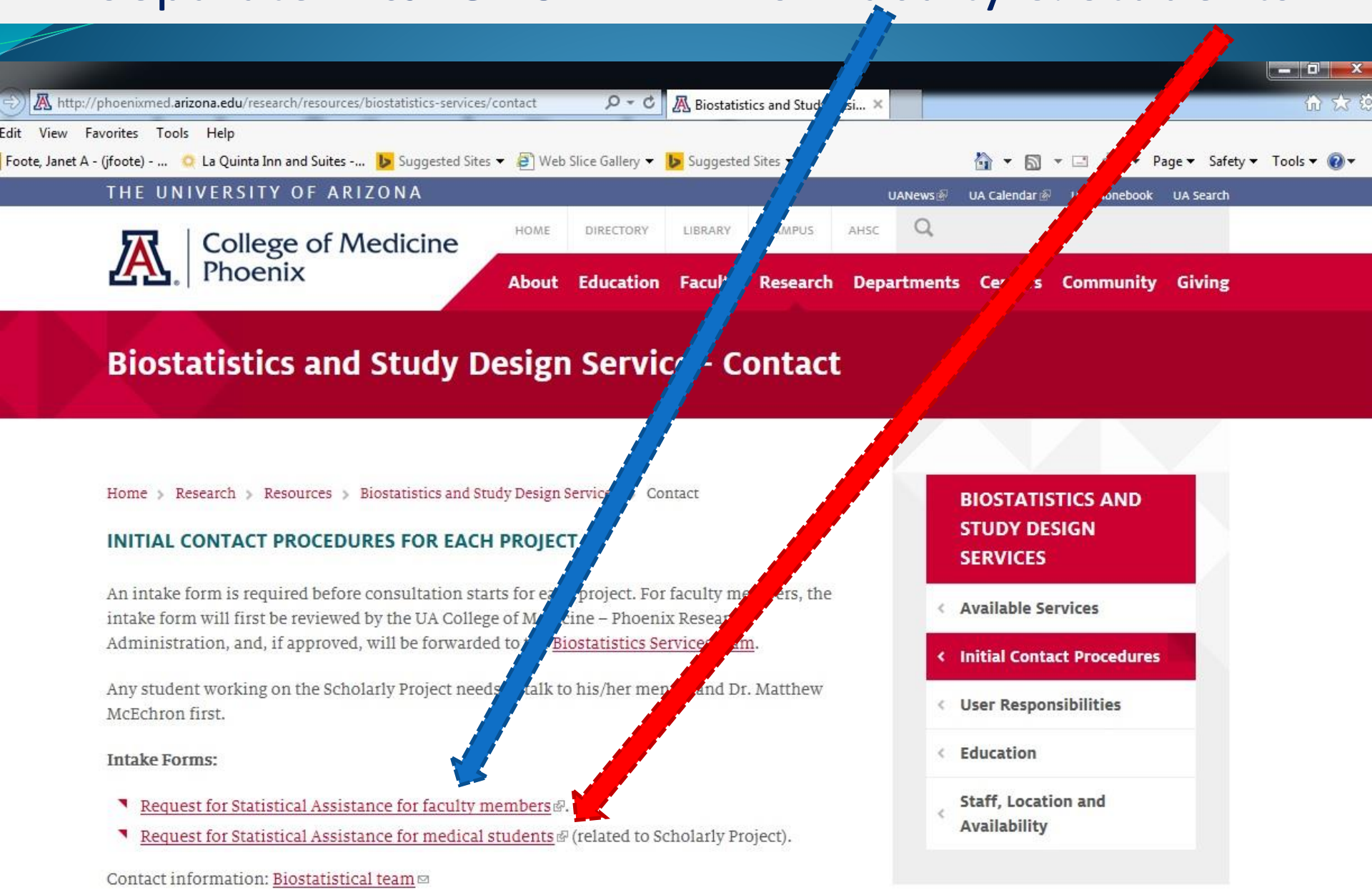


Biostatistics & Study Design Services page

http://phoenixmed.arizona.edu/research/resources/biostatistics-services



Separate **Intake Form** link for Faculty & Students



When should you contact Biostat support?

- As you are starting to design your study protocol
- Prior to submitting proposals for projects, routing
- Before you pilot your survey / start data collection
- As you are preparing to analyze data

Many roads…..

We are here to help !

What information do you need to have?

- What is your overall research question?
- What are your specific aims to answer your question?
- Who is you target population?
- What approach do you think you'll be using (i.e. measurements, surveys, observations/chart reviews, etc)?
- How much difference / change / variation is important?
- When do you need this?

Study Design 101

Research Question

In science, we take an organized, methodical approach to examine a question.

The first thing we need to do for research, is ask a question.

Once we have question, we need to complete background research. By reading what has been published about a topic, we often find out new information and sometimes change our question a bit because we are better informed about the topic.

The next important step in the scientific method is constructing a hypothesis.

A hypothesis is an 'educated idea' about how things will work.

There is a set way one to state the hypothesis.

Independent variable ***Dependent variable***

If _____*(I do this)*, then _____*(this)*_____ will happen.

Ex: *If I test the blood sugar of 100 adults, more than 20% will be at risk* for diabetes.*

If I add methyltrexate to HL1 cancer cells, the cells will stop multiplying, so the cancer cannot progress.

The hypothesis should be measureable, and do able.

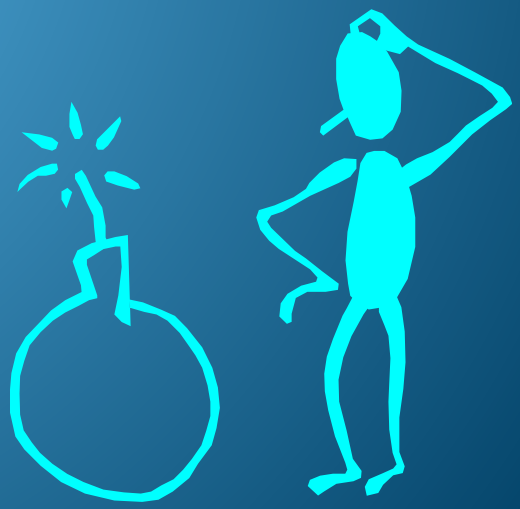
** Risk is defined as…….*

Once you have your hypothesis, you test it!

Testing a hypothesis is not a “one and done” procedure. In science, results must be shown to be repeatable and consistent. Statistics helps us understand the odds that the results we see are ‘real’ based on our study design.

Statistics:

- Help understand the odds that results are real
- Dependent on type and characteristics of data
- Cannot **fix** design / data / recruitment problems



Data: values, pieces of information

Types

- Categorical
- Quantity
- Nominal
- Ordinal
- Binary
- Discrete and continuous data.
- Interval and ratio variables
- Qualitative
- Quantitative

} Characteristics of data

Categorical Data

- The objects being studied are grouped into categories.
- Categories are usually based on a **qualitative** trait.
- These data are merely labels or categories.
- May or may not have any underlying order.

Nominal Data

Categorical data in which objects fall into ***unordered*** categories.

Examples:

- Type of Bicycle
  - Mountain bike, road bike, chopper, folding, BMX.
- Ethnicity
  - Asian, Pacific Islander, African American, Caucasian, Latino, Native American (note problems with these categories).
- Smoking status
  - smoker, non-smoker, former smoker

Ordinal Data

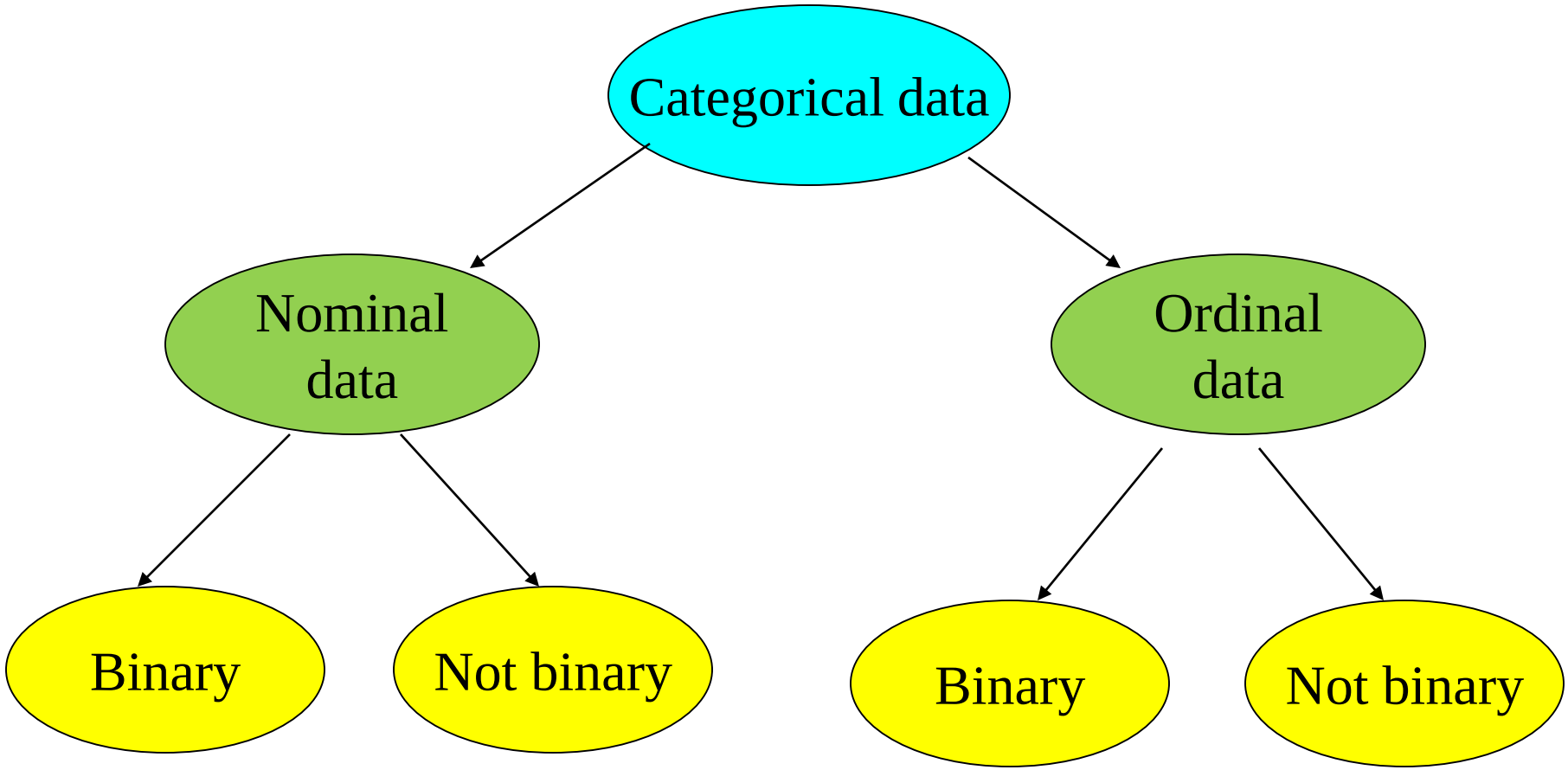
- Categorical data in which ***order*** is important.
- Highest Education level – elementary, high school, college graduate
- Degree of illness- none, mild, moderate, acute, chronic.
- Opinion of students about stats classes-

  Very unhappy, unhappy, neutral, happy, ecstatic!

Binary Data

- Special type of categorical data in which there are ***only two categories***.
- Binary data can either be nominal or ordinal.
- Current smoking status: smoker, non-smoker
- Attendance: present, absent
- Class mark: pass, fail.
- Status of student: undergraduate, postgraduate.

Categorical data classified as **Nominal**, **Ordinal**, and/or **Binary**



Quantity Data

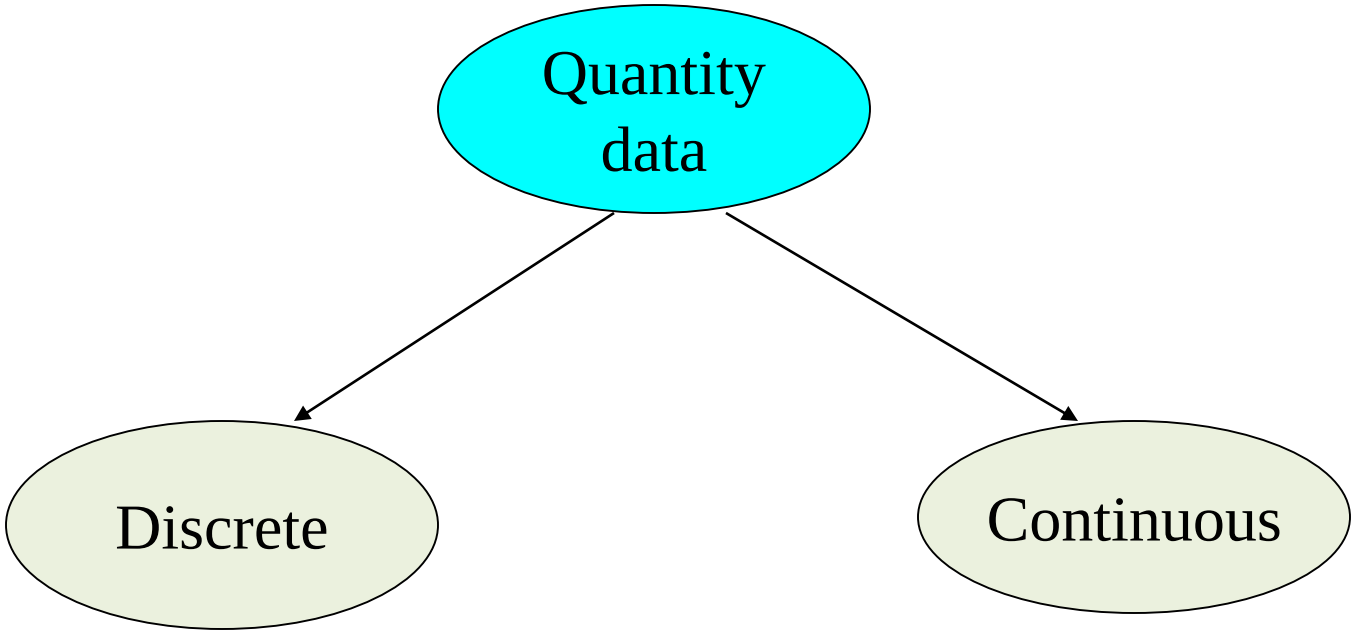
- Whatever is under study is being ‘measured’ based on some **quantitative** trait.
- Data are set of numbers.

Examples

- Pulse rate
- Height
- Age
- Exam marks
- Size of bicycle frame
- Time to complete a statistics test
- Number of cigarettes smoked

Quantity data can be classified as

Discrete or **Continuous**



Discrete Data

Only certain values are possible (there are gaps between the possible values). Implies counting.

Continuous Data

Theoretically, with a fine enough measuring device, no gaps.

Discrete Data

- Number of children in a family
- Number of students passing a stats exam
- Number of crimes reported to the police
- Number of bicycles sold in a day.

Generally, discrete data are counts.

We would not expect to find 2.2 children in a family or 88.5 students passing an exam or 127.2 crimes being reported to the police or half a bicycle being sold in one day.

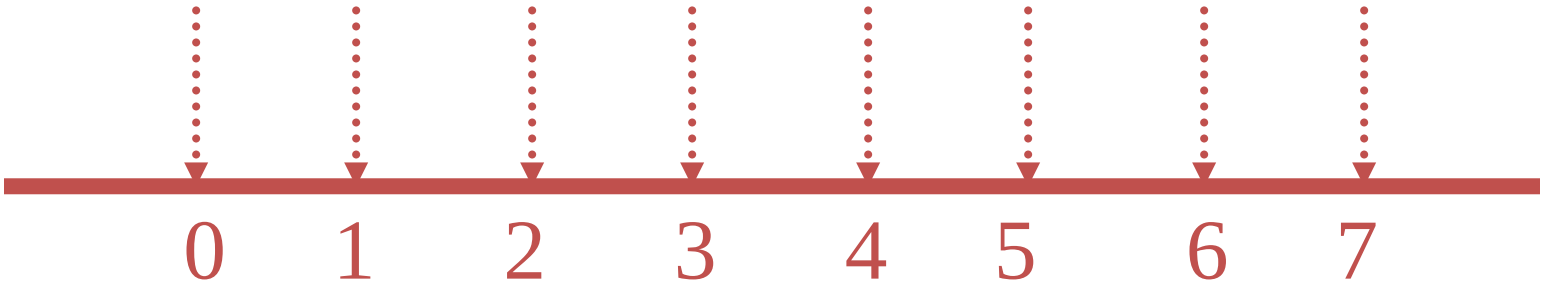
Continuous data

- Size of bicycle frame
- Height
- Time to run 500 metres
- Age

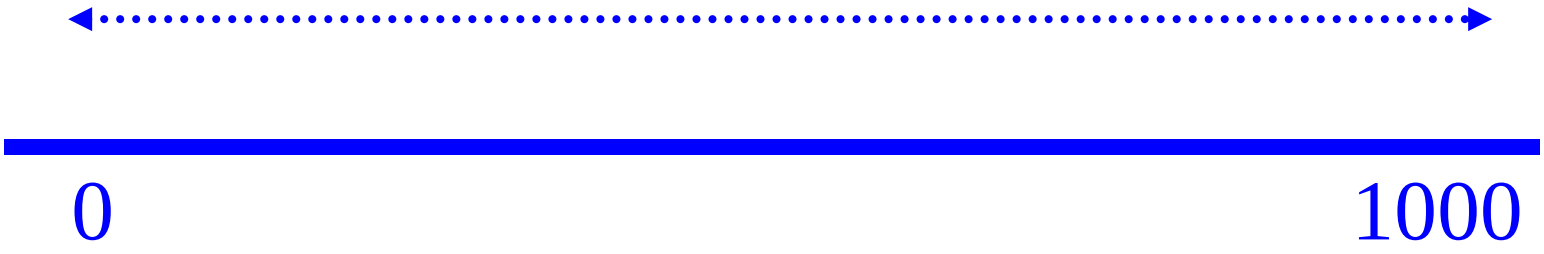
'Generally, continuous data come from measurements.

(any value within an interval is possible with a fine enough measuring device'- (Rowntree 2000)).

Discrete data -- Gaps between possible values- count



Continuous data -- *Theoretically*,
no gaps between possible values- measure



Why is this Important?

The type of data collected in a study determines the type of statistical analysis used.

What is a database?

A database is a **method of organizing and analyzing information.**

Why use a database?

- **Organize & analyze** information in different ways
  - Sorting
  - Grouping
  - Querying
  - Reporting
  - Exporting for statistical analysis
- Computerized database
  - Speed
  - Quality control
  - Precision
  - Automate repetitive tasks

Databases versus Excel

- Excel has some **limited** capabilities to sort data but its **primary function is to create financial spreadsheets**
  - Can create "what if" scenarios to determine financial consequences
  - **Can be used for small /limited research data sets & simple lists**
  - Not multi-user such that only one person can work on the file at a time
- Databases: **designed to collect, sort, & manipulate data**
  - Databases can process large amounts of data; usually limited by hardware constraints
  - Structure is in the same format for each member record of a table
  - Data quality control features ensure that valid data is entered
  - A relational database allows for linking of an unlimited number of tables
  - Databases are multi-user because the data can reside on a server and multiple people can have access at the same time
  - Many databases offer web interfaces thereby eliminating the need for each user to have a copy of the program on their computer

Databases versus Excel

- Many databases offer audit functions required by certain regulatory agencies
  - Tracks date record created and modified
  - Tracks original and changed values
  - Requires user to give reason for the change
- Databases are more suitable for importing data from multiple sources
  - More robust in connecting to different data sources
  - Imports of different data types into different tables can be linked via common identifiers such as subject ID
  - Merging multiple data sources into Excel so that the rows line up properly in a flat file format can be a challenge

U of A resources

- Qualtrics™ – Survey Monkey on steroids
- Redcap™ – Clinical database

How is a database organized?

- One or more tables
- Tables store records
  - Patient identifiers
  - Demographics and history
  - Test results
  - Etc…..
- A record is a collection of fields
  - Patient identifiers
    - Name, DOB, address, …..are stored in separate fields

Records and Fields

Fields

Records

ID	Age	Gender	Group	Race	Sex
3001	50.48	Male	Combined	CC	0
3002	65.55	Male	Diet	AA	0
3003	63.59	Female	Diet	CC	1
3005	50.07	Female	Combined	CC	1
3010	60.28	Male	Diet	CC	0
3011	56.43	Female	Diet	CC	1
3012	45.80	Female	Combined	CC	1
3013	56.05	Female	Combined	CC	1
3014	65.48	Female	Diet	AA	1
3015	58.21	Female	Diet	CC	1
3016	57.30	Female	Combined	CC	1
3017	53.93	Female	Combined	CC	1
3018	50.12	Female	Diet	CC	1
3019	57.36	Female	Combined	CC	1
3020	51.05	Male	Diet	CC	0
3021	66.11	Female	Diet	CC	1
3024	54.90	Female	Diet	AA	1
3025	65.62	Female	Combined	CC	1
3027	45.91	Female	Diet	AA	1
3029	58.42	Female	Combined	CC	1
3032	53.50	Male	Diet	CC	0

How is data displayed?

- Fields are displayed on layouts
  - Forms
  - Web
  - Reports
- Data can be from a single table or many tables if using a relational database

Relational Database Example

Subject Info

Id	Name	Age
10	Smith	50
11	Jones	55
12	Doe	60

Anthropometrics

ID	Weight (lb)	Weight (kg)
10	230	104.5
11	212	96.4
12	199	90.4

Physical Activity

ID	KCAL	KCAL/kg
10	2400	23.1
11	2652	27.5
12	2350	25.9

Treadmill Performance

ID	V02	V02/kg
10	2.8	26.7
11	3.2	33.1
12	2.1	23.2

Differences between a clinical & research database

- Clinical database
  - Form or report oriented so data is displayed for clinical decision making
  - Emphasis on displaying or reporting of individual data rather than accumulating multiple records
- Research database
  - Table oriented so that data is accumulated for eventual export to a statistical package for data analysis and reporting
  - Less emphasis on individual records

Clinical data:

SHAPE II / III Metabolic Stress Test

Subject ID	2002
Sequence	◉ Baseline ○ 6 mo
Visit Date	06/16/2004

Weight (lb)	185.0
HR Rest (bpm)	67.
Time (max sec)	534
Max Speed (mph)	3.0
Max Elev (%)	5.0
Max VO2/kg	21.1
Max VO2 (L/min)	1778
Max VCO2 (L/min)	1992
Max ReR	1.12
Maximal HR (bpm)	134
Max VO2/HR	13
Max RR	48
Max VE (L/min)	83.7
Max RPE	9.
Max SBP (mmHg)	220.
Max DBP (mmHg)	88.

AT

% VO2 Max	61.00
HR (bpm)	112
VO2/kg (mL/kg/min)	13.00

Research data:

ID1	Seq	VisitDate	WeightLb	HRRest	Time	Speed	Elev	VO2kg	VO2	VC02	RER
2002	◉ Baseline ○ 6 mo	06/16/2004	185.0	67.	534	3.0	5.0	21.1	1778	1992	
2001	◉ Baseline ○ 6 mo	06/09/2004	195.0	48.	746	03.0	10.0	18.5	1637	1763	
2004	◉ Baseline ○ 6 mo	06/10/2004	239.0	91.	854	03.0	10.0	19.2	2081	2051	
2005	◉ Baseline ○ 6 mo	06/18/2004	280.0	97.	514	3.0	5.0	20.6	2620	2732	
2006	◉ Baseline ○ 6 mo	07/15/2004	204.0	98.	660	3.0	7.5	18.3	1697	1819	
2011	◉ Baseline ○ 6 mo	07/15/2004	169.0	75.	557	3.0	7.0	11.7	0898	0908	
2020	◉ Baseline ○ 6 mo	09/01/2004	243.4	63.	444	3.0	5.0	13.7	1512	1691	
2025	◉ Baseline ○ 6 mo	08/24/2004	200.0	72.	820	3.0	10.0	22.9	2085	2357	
2019	◉ Baseline ○ 6 mo	09/07/2004	133.9	63.	857	3.0	10.0	21.5	1307	1466	
2028	◉ Baseline ○ 6 mo	10/07/2004	222.0	85.	1134	3.0	15.0	28.7	2895	3035	
2017	◉ Baseline ○ 6 mo	10/15/2004	211.0	66.	540	3.0	5.0	17.9	1715	1809	
2041	◉ Baseline ○ 6 mo	10/25/2004	216.0	83.	629	3.0	7.5	18.4	1803	1783	
2035	◉ Baseline ○ 6 mo	11/04/2004	209.0	92.	783	3.0	10.0	23.4	2233	2260	
2040	◉ Baseline ○ 6 mo	11/04/2004	284.0	88.	360	3.0	2.5	20.0	2585	2613	
2029	◉ Baseline ○ 6 mo	09/16/2004	168.0	85.	777	3.0	10.0	18.8	1433	1467	
2016	◉ Baseline ○ 6 mo	08/02/2004	183.0	69.	869	3.0	10.0	23.1	1918	2084	
2034	◉ Baseline ○ 6 mo	11/18/2004	196.0	85.	1168	3.0	15.0	31.1	2773	3280	
2045	◉ Baseline ○ 6 mo	11/23/2004	245.9	69.	813	3.0	10.0	23.1	2577	2700	
2039	◉ Baseline ○ 6 mo	12/03/2004	250.0	78.	474	3.0	5.0	14.8	1683	1906	
2051	◉ Baseline ○ 6 mo	12/06/2004	215.0	68.	670	3.0	7.5	17.7	1728	1740	
2005	○ Baseline ◉ 6 mo	12/28/2004	280.0	97.	638	3.0	7.5	22.0	2806	2813	
2056	◉ Baseline ○ 6 mo	01/18/2005	140.5	59.	744	2.0	10.0	14.6	0931	0924	
2025	○ Baseline ◉ 6 mo	01/31/2005	200.0	85.	694	3.0	7.5	21.0	1912	2115	
2053	◉ Baseline ○ 6 mo	02/08/2005	204.5	91.	473	3.0	5.0	16.5	1536	1566	
2016	○ Baseline ◉ 6 mo	02/04/2005	180.0	73.	814	3.0	10.0	24.0	1964	2140	
2004	○ Baseline ◉ 6 mo	02/14/2005	252.0	72.	885	3.0	10.0	21.4	2445	2515	
2001	○ Baseline ◉ 6 mo	02/14/2005	195.0	39.	812	3.0	10.0	17.8	1580	1733	
2006	○ Baseline ◉ 6 mo	01/31/2005	202.0	111.	649	3.0	7.5	18.3	1679	1700	
2061	◉ Baseline ○ 6 mo	02/16/2005	226.0	88.	913	3.0	10.0	25.3	2603	2603	
2052	◉ Baseline ○ 6 mo	03/04/2005	203.0	72.	1066	3.0	12.5	27.2	2506	2687	
2066	◉ Baseline ○ 6 mo	03/16/2005	257.0	87.	982	3.0	12.5	24.1	2816	3063	
2069	◉ Baseline ○ 6 mo	03/16/2005	200.2	55.	616	3.0	7.5	18.1	1653	1733	
2063	◉ Baseline ○ 6 mo	03/22/2005	260.0	59.	583	3.0	7.5	16.9	2002	2076	

Advantages of a database

- Collection of data in a centralized location
- Controls redundant data
- Data stored so as to appear to users in one location
  - Data can be stored in multiple tables and come from multiple sources
  - A relational database brings it all together

Sharing and Exchanging Data

- Multiple users can access the same database via a network
  - Can be local or over the internet
  - Best done when the data are stored on a database server
    - Access via a client application
    - Access via a web interface
  - Server allows remote access over the internet from anywhere
    - Should be behind a firewall for security with access via VPN and password protection

Database Design Considerations

- **What to collect**
  - What questions are to be answered?
  - Think of the data tables in your future publications
    - **Focus on the key data elements rather than collect as much as possible**
- What statistical package will be used?
  - Format of the data file to which the data will be exported
    - Allowable characters
    - Format for certain analyses
      - For example, gender can be recorded in the database as M or F but statistical package may require 0 and 1
    - Length of data field labels
    - Long or wide format

Long versus Wide Format

Long: each year is represented as its own observation in a record

	famid	year	faminc
1.	1	96	40000
2.	1	97	40500
3.	1	98	41000
4.	2	96	45000
5.	2	97	45400
6.	2	98	45800
7.	3	96	75000
8.	3	97	76000
9.	3	98	77000

Wide: each family is a record and each year is a field with that record

	famid	faminc96	faminc97	faminc98
1.	3	75000	76000	77000
2.	1	40000	40500	41000
3.	2	45000	45400	45800

Selected Elements of Data Management Planning

Quality Control of Data Before Study

- Collect only needed variables
- Select appropriate computer hardware & software
- Plan analyses with dummy tabulations
- Develop study forms
  - Precode responses
  - Format boxes for data entry
  - Label each page with date, time, ID
  - Consider scan technology

What needs to be in the research database?

- Research variables directly related to the hypotheses being tested-**YES**
- Clinical measures used for screening-**MAYBE**
  - Blood work, ECG, medical history
- Administrative data-**NO**
  - Contact information
  - Scheduling

Where Are the Original Data?

In the source documents

What is a Source Document?

- **It is the First Recording**

- **What does it tell?**

    1. It is the data that document the trial

    2. Study was carried out according to protocol

Source Documents

- Original Lab reports
- Pathology reports
- Surgical reports
- Physician Progress Notes
- Nurses Notes
- Medical Record
- Letters from referring physicians
- Original radiological films
- Tumor measurements
- Patient Diary/patient interview

Common Data Elements

- **Standardized**, unique terms and phrases that delineate discrete pieces of information used to collect data in a clinical trial
- **Uniform** representation of demographics and data points to consistently track trends
- Elements **define** study parameters and **endpoints**

Designing the questions

- Granular primary data
  - No observer conclusions, synthesis, coding
- Categorical/ordinal data when possible—statistical power. Re-slice at analysis
- Use validated scales/instruments
  - Don’t build your own unless unavoidable
- Collect key variables with >1 question
- Avoid measurements that cluster at one end of scale
  - Distribution problems, Likert scales

Forms Design

CDC

Influenza-Associated Pediatric Deaths Case Report Form

Form approved
OMB No. 0920-0007

STATE USE ONLY – DO NOT SEND INFORMATION *IN THIS SECTION* TO CDC

Last Name: _____ First Name: _____ County: _____

Address: City: State, Zip:

Patient Demographics

1. State:	2. County:	3. State ID:	4. CDC ID:
5. Age: ____ ☐ Days ☐ Months ☐ Years	6. Date of birth: ____/____/____ MM DD YYYY	7.Sex: ☐ Male ☐ Female	8. Ethnicity: ☐ Hispanic or Latino ☐ Not Hispanic or Latino ☐ Unknown

9. Race: ☐ White ☐ Black ☐ Asian ☐ Native Hawaiian or Other Pacific Islander ☐ American Indian or Alaska Native ☐ Unknown

Death Information

10. Date of illness onset: ____/____/____ MM DD YYYY	11. Date of death: ____/____/____ MM DD YYYY	12. Was an autopsy performed? ☐ Yes ☐ No

13. Location of death: ☐ Home ☐ Emergency Dept (ER) ☐ Inpatient ward ☐ ICU ☐ Other (specify): _____

Influenza Testing (check all that were used)

Test Type	Result	Specimen Collection Date
☐ Commercial rapid diagnostic test	☐ Influenza A ☐ Influenza B ☐ Negative ☐ Influenza A/B (Not Distinguished)	____/____/____
☐ Viral culture	☐ Influenza A (Subtyping Not Done) ☐ Influenza B ☐ Negative ☐ Influenza A (Unable To Subtype) ☐ Influenza A (H1) ☐ Influenza A (H3)	____/____/____
☐ Direct fluorescent antibody (DFA)	☐ Influenza A ☐ Influenza B ☐ Negative ☐ Influenza A/B	____/____/____
☐ Indirect fluorescent antibody (IFA)	☐ Influenza A ☐ Influenza B ☐ Negative ☐ Influenza A/B	____/____/____
☐ Enzyme immunoassay (EIA)	☐ Influenza A (Subtyping Not Done) ☐ Influenza B ☐ Negative ☐ Influenza A (Unable To Subtype) ☐ Influenza A (H1) ☐ Influenza A (H3)	____/____/____
☐ RT-PCR	☐ Influenza A (Subtyping Not Done) ☐ Influenza B ☐ Negative ☐ Influenza A (Unable To Subtype) ☐ Influenza A (H1) ☐ Influenza A (H3)	____/____/____
☐ Immunohistochemistry (IHC)	☐ Influenza A ☐ Influenza B ☐ Negative	____/____/____

Culture confirmation of INVASIVE bacterial pathogens

14. Was an INVASIVE bacterial infection confirmed by culturing an organism from a specimen collected from a normally sterile site (e.g., blood, cerebrospinal fluid [CSF], tissue, or pleural fluid)? ☐ Yes ☐ No

☐ *Streptococcus pneumoniae*	☐ *Staphylococcus aureus*, methicillin **sensitive**	☐ *Neisseria meningitidis (serogroup, if known)*:_____
☐ *Haemophilus influenzae* type b	☐ *Staphylococcus aureus*, methicillin **resistant (MRSA)**	☐ Group A streptococcus
☐ *Haemophilus influenzae* not-type b	☐ *Staphylococcus aureus*, **sensitivity not done**	☐ Other invasive bacteria: _____

Operations Manual

- Defines entire study protocol, sequence
- Form-specific annotation, guidance
- Documents all post-hoc validity checks, edit checks, data curation criteria
- Evolving document with periodic updates
  - Preferably on-line
- Use for training, quality control, process planning

Data Dictionary - Operational

- For every form/table, lists:
  - Variable name (database field)
  - Variable description (plain English)
  - Variable type (string, integer, numeric, etc.)
  - Variable length (or precision)
  - Nullability (missing or no value indicator)
  - Range checks, allowable values
  - Coding conventions, with definitions

Data Dictionary

Variable name	Code	Description
ANYSKCA	1=yes; 0=no	Any NMSC post-randomization?
ANYSKCA6	1=yes; 0=no	Any NMSC after 6 months post-randomization?
BCCOCC	1=yes; 0=no	Any BCC occurrence post-randomization?
BCCOCC6	1=yes; 0=no	Any BCC after 6 months post-randomization?
SCCOCC	1=yes; 0=no	Any SCC occurrence post-randomization?
SCCOCC6	1=yes; 0=no	Any SCC after 6 months post-randomization?
ALLSKCA	number	Total number of NMSC which occurred post-randomization
ALLSKCA6	number	Total number of NMSC after 6 months post-randomization
ALLBCC	number	Total number of BCC which occurred post-randomization
ALLBCC6	number	Total number of BCC after 6 months post-randomization
ALLSCC	number	Total number of SCC which occurred post-randomization
ALLSCC6	number	Total number of SCC after 6 months post-randomization
ANYMOS	number	Total number of months before any NMSC occurrence
ANYMOS6	number	Months (after 6 mos post-randomization) before NMSC
BCCMOS	number	Total number of months before first BCC occurrence
BCCMOS6	number	Months (after 6 mos post-randomization) before BCC

Why code:

- Forces analyzable data structure, format
- Vastly simplifies analysis
- Speeds data input/transcription
- Vastly simplifies data analysis/reporting

Example of the need for data coding

What is the subject's sex?

male	female
Male	Female
M	F
m	f
Man	Woman
Boy	Girl
0	1
1	2
Gentleman	Lady
Tarzan	Jane

What do you mean & how will you record it?

- HEADACHE
  - Headache
  - Pain in the head
- ACHE:
  - Ache:Head
  - Head Pain
  - HP

Unless there is a standard code for the use of terms, data retrieval becomes difficult

Rules for Data Entry

- Each variable has a field in the dataset
- Categorical and nominal values require a number or string code
- Continuous values are entered directly
- Missing values must be different values from a real response
  - Common formats are “99” or bullets “·”
  - Don’t know is a response—do not leave blank
  - “0” is not the same as missing
- Coding instructions should be on form
- Avoid open-ended questions

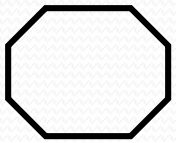
Avoid open-ended questions

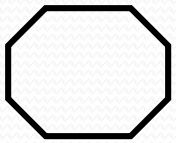
Enter the subject’s gender:_____

Enter the subject's level of education:_____

Close Ended Question

What is the subject's sex? *Check one*

 Male

 Female

Use pre-coded responses where possible

Subject ID 1001

Gender ◉ Male ○ Female

Age 56

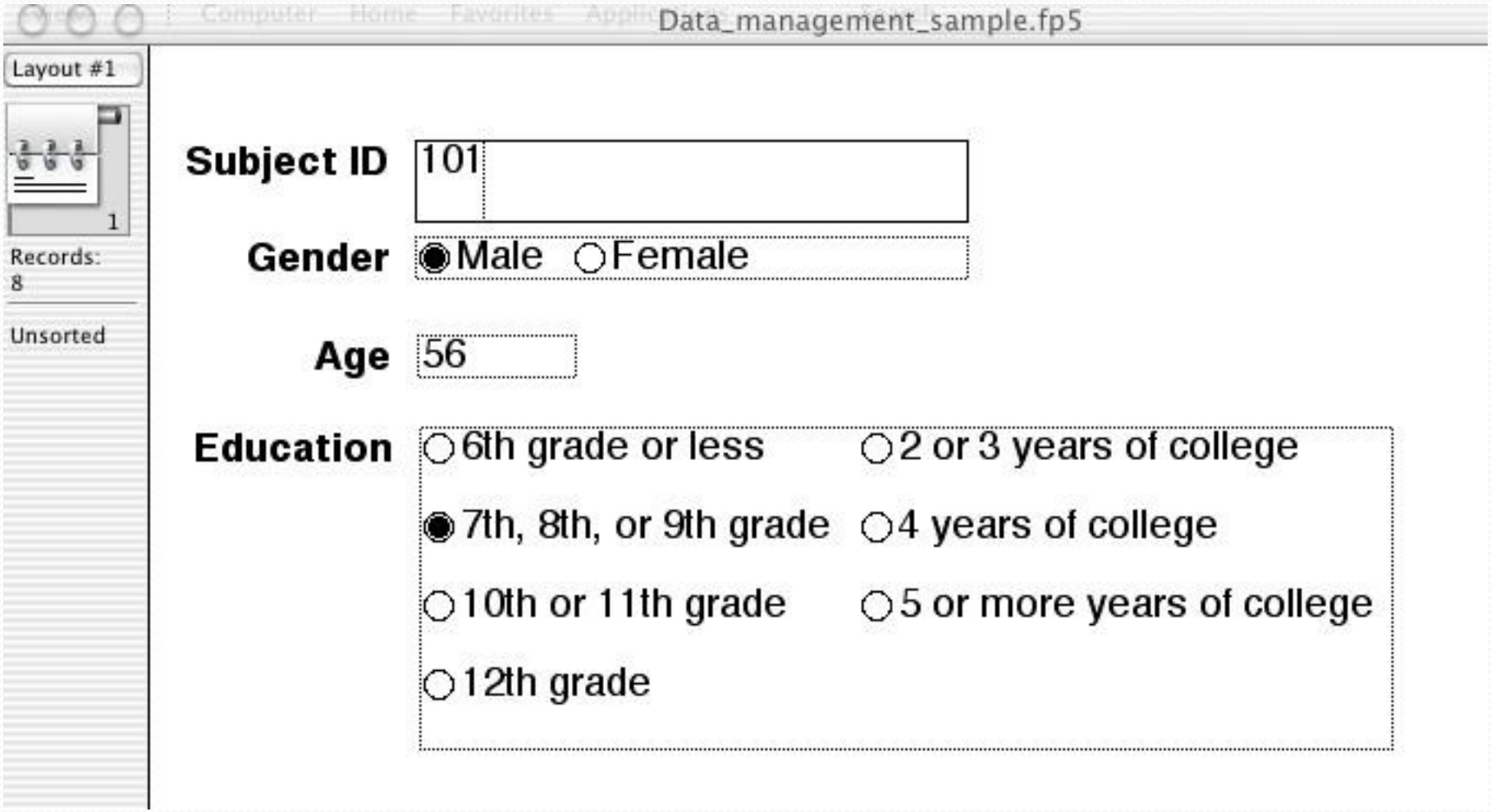
Education

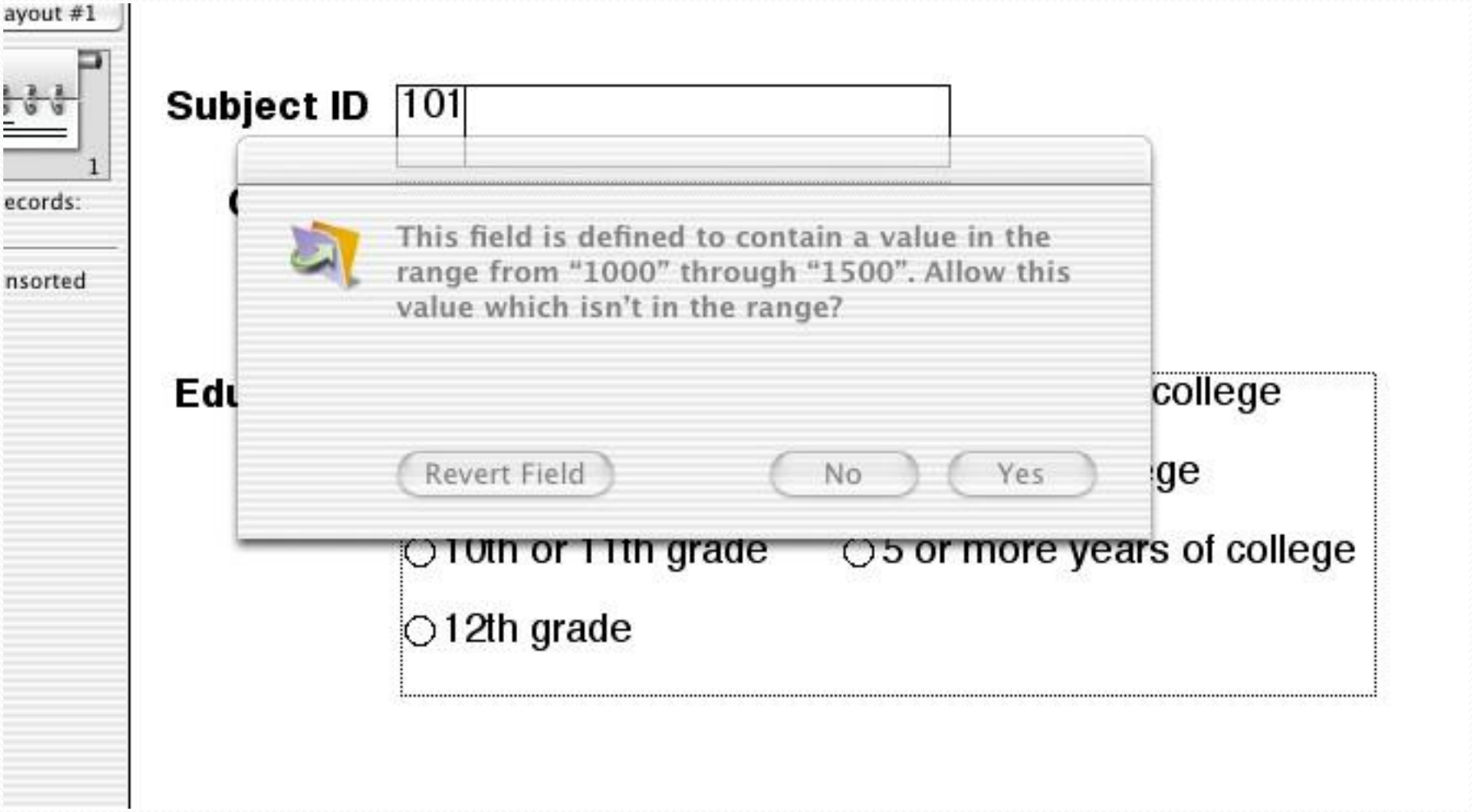
○ 6th grade or less	○ 2 or 3 years of college
◉ 7th, 8th, or 9th grade	○ 4 years of college
○ 10th or 11th grade	○ 5 or more years of college
○ 12th grade	

Data in Spreadsheet

Subject ID	Gender	Age
1001	Male	52
1002	Male	54
103	Mael	65
1004	Female	54
5	Female	52
1006	Female	52
1007	Femele	75
1008	Male	48
1009	M	37
1010	Female	73
11	F	54

Data Validation


Data_management_sample.fp5
Layout #1
1
Records:
8
Unsorted
Subject ID
101
Gender
Male
Female
Age
56
Education
6th grade or less
7th, 8th, or 9th grade
10th or 11th grade
12th grade
2 or 3 years of college
4 years of college
5 or more years of college


ayout #1
1
ecords:
nsorted
Subject ID
101
This field is defined to contain a value in the range from “1000” through “1500”. Allow this value which isn’t in the range?
Revert Field
No
Yes
Edu
college
ge
10th or 11th grade
5 or more years of college
12th grade

Types of Edit Checks

- Patient identification and record linkage
  - ID #’s, name spelling, ID#’s on all pages
- Legibility
- Correct form for examination
- Missing data
- Consistency
- Range and inadmissible codes

Backup

- Data must be backed up on a regular basis to protect against:
  - Theft, fire, floods, hurricanes,
  - Equipment failure
- Computer backup
  - Mirrored drives
  - Digital tapes
  - Store backup tapes off-site

Putting it All Together: Research Data Management

- An artful selection of physical & electronic management methods
  - Signed informed consent documents
  - Paper forms
  - Regulatory & project management binders
  - Data models and databases
  - Data acquisition and display technologies
  - Communications technologies for project management as well as data management

Attributes of Successful Data Management

- Attention to detail
- Explicit structure and process
- Robust designs
  - Anticipate failures, lapses and mistakes
  - Design systems that identify and correct them
- Mechanisms for verification
- Well documented

Quality

Fast is fine, but accuracy is everything.

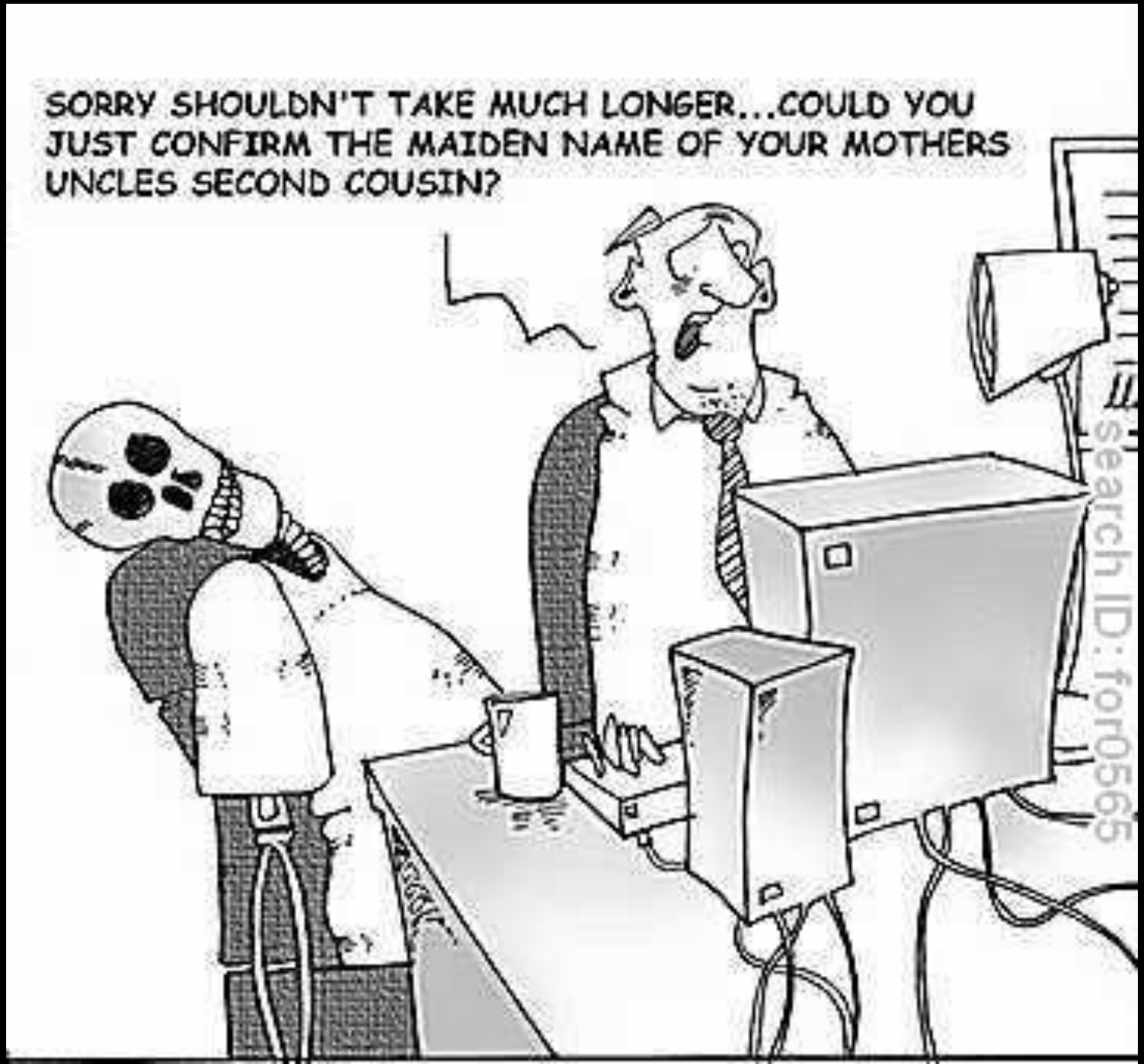
(Wyatt Earp)

Study Designs

Study Design	Study Design Decision Short-hand	Measure of Association	Interpretation of Measure of Association
Case control	**Starts with cases**	**Odds ratio (OR)**	**Those with the disease were X times more/less likely to have been exposed**
Cohort	**Starts with a designated group or exposure; follows through time**	**Relative risk (RR)**	**Those with the exposure were X times more/less likely to have the disease**
Clinical Trial	**Investigator decides exposures**		
Cross-sectional	**None of the above**		

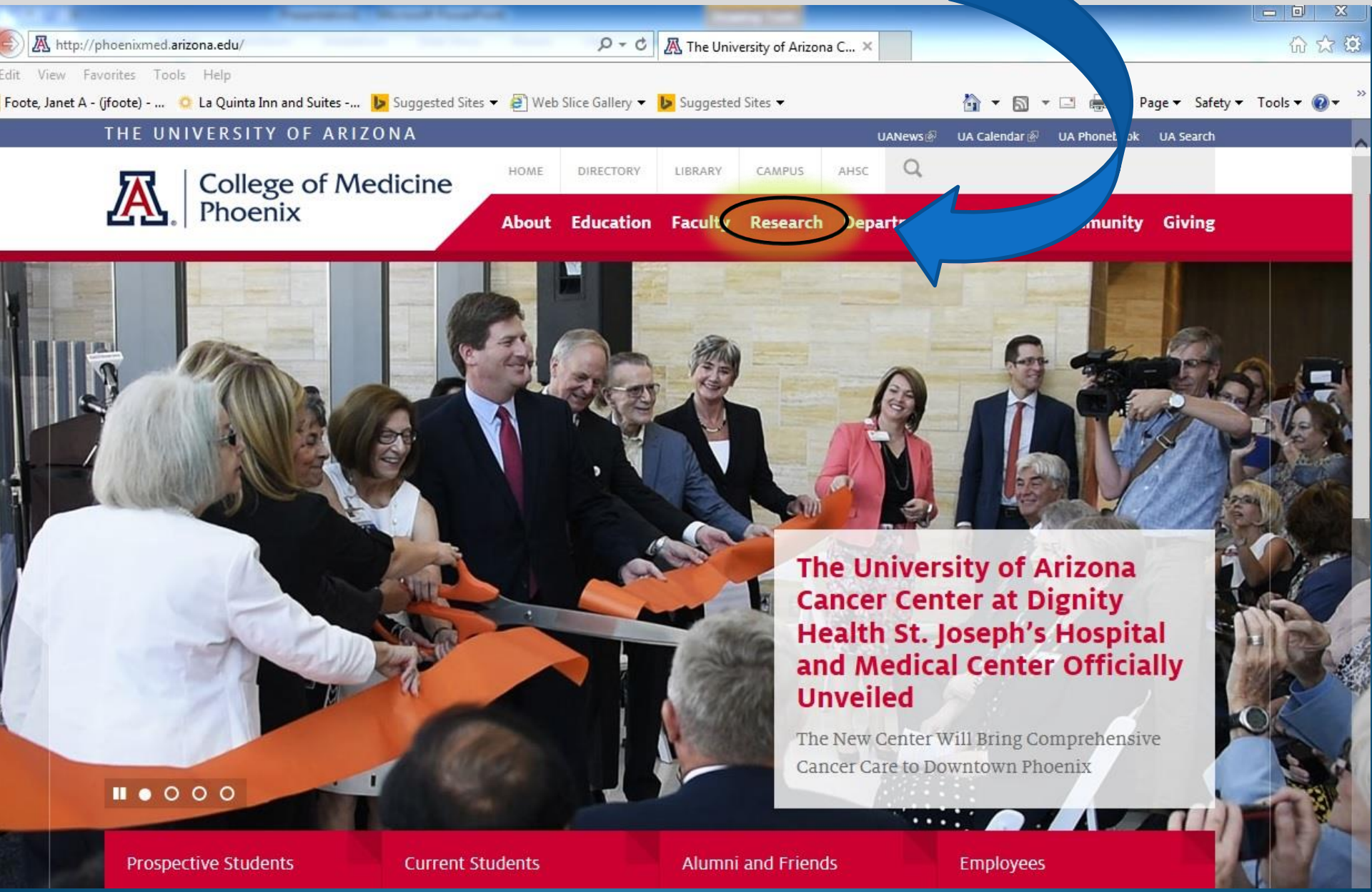
Want to be thorough…..

…but as minimally-invasive as possible

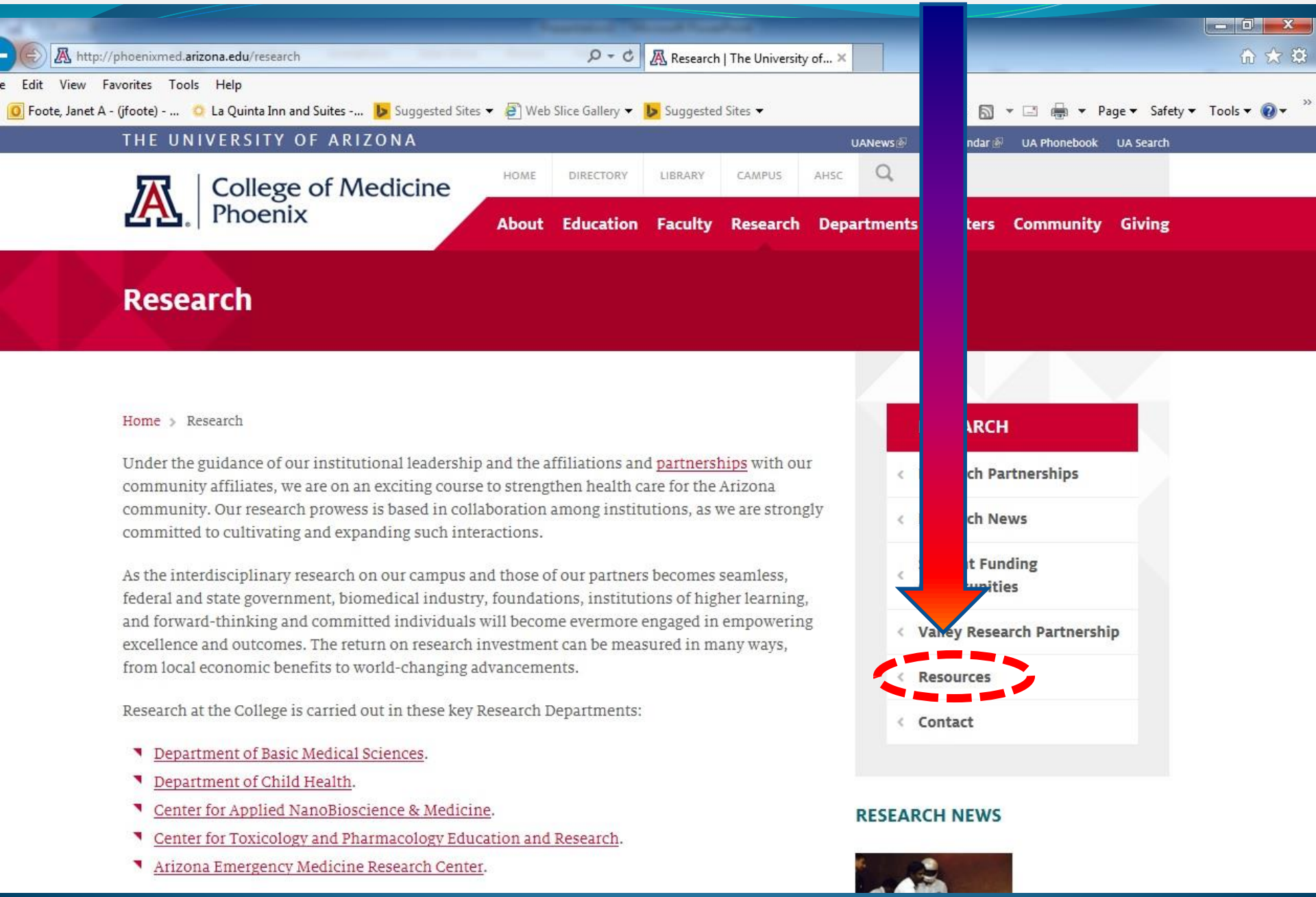

SORRY SHOULDN'T TAKE MUCH LONGER...COULD YOU JUST CONFIRM THE MAIDEN NAME OF YOUR MOTHERS UNCLES SECOND COUSIN?
search ID: for0565

Finding more information online

Research

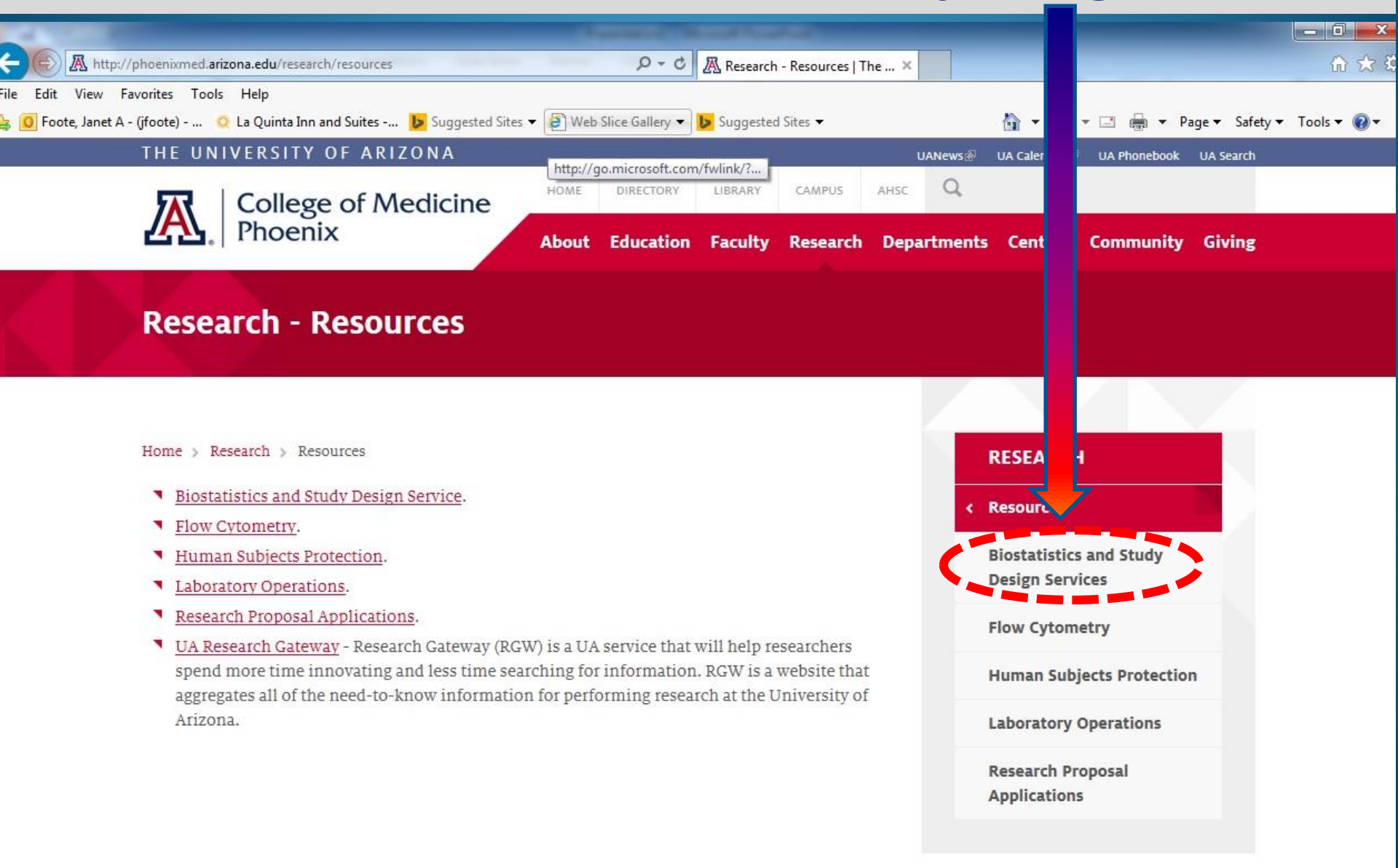


On the Research page, select **Resources**



Under Resources, select

Biostatistics & Study Design Services



Biostatistics & Study Design Services page

http://phoenixmed.arizona.edu/research/resources/biostatistics-services