

Minor Detail #1:

What's the BCC?

BCC: Biostatistics Collaboration Center

Who We Are



Leah J. Welty, PhD
Assoc. Professor
BCC Director



Lauren Balmert, PhD
Asst. Professor



Jody D. Ciolino, PhD
Asst. Professor



Kwang-Youn A. Kim, PhD
Assoc. Professor



Masha Kocherginsky, PhD
Assoc. Professor



Mary J. Kwasny, ScD
Assoc. Professor



Julia Lee, PhD, MPH
Assoc. Professor



David Aaby, MS
Senior Stat. Analyst



Elizabeth Gray, MS
Stat. Analyst



Kimberly Koloms, MS
Stat. Analyst



Amy Yang, MS
Senior Stat. Analyst



Tameka L. Brannon
Financial | Research
Administrator

BCC: Biostatistics Collaboration Center

Our Mission

Mission: to support investigators in the conduct of high-quality, innovative health-related research by providing expertise in biostatistics, statistical programming, and data management.

How do we accomplish this?

1. Every investigator is provided a **FREE** initial consultation of 1-2 hours, subsidized by **FSM Office for Research**. Thereafter:
 - a) Grants
 - b) Subscription
 - c) Re-charge (Hourly) Rates
2. Grant writing (e.g. developing analysis plans, power/sample size calculations) is also supported by FSM at **no cost to the investigator**, with the goal of establishing successful collaborations.

BCC: Biostatistics Collaboration Center

What We Do

Many areas of expertise, including:

- Bayesian Methods
- Big Data
- Bioinformatics
- Causal Inference
- Clinical Trials
- Database Design
- Genomics
- Longitudinal Data
- Missing Data
- Reproducibility
- Survival Analysis

Many types of software, including:



BCC: Biostatistics Collaboration Center

An Overview of Shared Statistical Resources



Biostatistics Collaboration Center (BCC)

- Supports non-cancer research at NU
- Provides investigators an initial 1-2 hour consultation subsidized by the FSM Office of Research
- Grant, Hourly, Subscription



Quantitative Data Sciences Core (QDSC)

- Supports all cancer-related research at NU
- Provides free support to all Cancer Center members subsidized by RHLCCC
- Grant

Biostatistics Research Core (BRC)

- Supports Lurie Children's Hospital affiliates
- Provides investigators statistical support subsidized by the Stanley Manne Research Institute at Lurie Children's.
- Hourly

BCC: Biostatistics Collaboration Center

Shared Resources Contact Info

- Biostatistics Collaboration Center (BCC)
 - Website: <http://www.feinberg.northwestern.edu/sites/bcc/index.html>
 - Email: bcc@northwestern.edu
 - Phone: 312.503.2288
- Quantitative Data Sciences Core (QDSC)
 - Website: http://cancer.northwestern.edu/research/shared_resources/quantitative_data_sciences/index.cfm
 - Email: qdsc_rhlccc@northwestern.edu
 - Phone: 312.503.2288
- Biostatistics Research Core (BRC)
 - Website: <https://www.luriechildrens.org/en-us/research/facilities/Pages/biostatistics.aspx>
 - Email: merreed@luriechildrens.org
 - Phone: 773.755.6328

Minor Detail #2:

Assuming observations are
independent

Independent Observations: Overview

- Many common statistical methods assume observations are independent (nearly everything taught in a usual statistics course)
- There are different statistical methods for observations that are not independent
- Examples of paired/not independent data
 - Before and after measurements
 - Case and matched control
 - Longitudinal data
 - Nested samples
 - Spatial data
- Analyses that assume observations are independent, when in reality they're not, can be very wrong

(In)dependence Example: Two Case-Control Studies

Hodgkins & Tonsillectomy

- Is Tonsillectomy associated with Hodgkin's?
- Vianna, Greenwald, and Davies (1971)
 - Case-control study (controls unmatched)
- Johnson & Johnson (1972)
 - Case-control study (controls matched)

Adapted from Mathematical Statistics and Data Analysis, John A. Rice, Duxbury (1995)

(In)dependence: Contingency Table Vianna et al.

Vianna et al.

	Tonsillectomy	No Tonsillectomy
Hodgkin's (n = 101)	67	34
Control (n = 107)	43	64

- Case-control study
 - Recruit people with Hodgkin's and similar people without
- Look back to see who had exposure (tonsillectomy)
 - In Hodgkin's group, $67/101 = 66\%$
 - In Control group, $43/107 = 40\%$
- Is that a big enough difference to conclude that tonsils are protective?

(In)dependence: Odds and Odds Ratios

Vianna et al.

	Tonsillectomy	No Tonsillectomy
Hodgkin's (n = 101)	67	34
Control (n = 107)	43	64

- Odds of tonsillectomy in Hodgkin's group: 67/34
- Odds of tonsillectomy in Control group: 43/64
- Odds ratio comparing tonsillectomy for Hodgkin's versus Control
 - $OR = (67/34)/(43/64) = 2.93$
 - "Hodgkin's had 2.93 times the odds of tonsillectomy compared to Controls."
- Odds ratios range from 0 to ∞
 - 1 = no difference in groups
- Is 2.93 different enough from 1 to conclude that tonsils are protective?

(In)dependence: Chi-Squared Test

Vianna et al.

	Tonsillectomy	No Tonsillectomy
Hodgkin's (n = 101)	67	34
Control (n = 107)	43	64

- A chi-squared test can be used to compare whether rows and columns in a 2x2 contingency table are associated
- Computed by comparing “expected” versus observed values
 - E.g. Expect 53.4 people to have Hodgkin's and a Tonsillectomy, observe 67
 - $101 * (67+43)/208$
- Chi-squared statistics is 14.46 with 1 degree of freedom
- P-value = 0.0002
- Conclude there is evidence for an association between Hodgkin's and Tonsillectomy

(In)dependence: A second study, Johnson et al.

Johnson et al.

	Tonsillectomy	No Tonsillectomy
Hodgkin's (n = 85)	41	44
Control (n = 85)	33	52

- Case-control study (**controls matched**)
 - 85 Hodgkin's who had sibling w/in 5 yrs age and same sex
 - Sibling was *matched* control

(In)dependence: What went wrong?

Johnson et al. NEJM

	Tonsillectomy	No Tonsillectomy
Hodgkin's (n = 85)	41	44
Control (n = 85)	33	52

- Look back to see who had exposure (Tonsillectomy)
 - In Hodgkin's group, $41/85 = 48\%$
 - In Control group, $33/85 = 39\%$
- Odds of tonsillectomy
 - In Hodgkin's group, $41/44$
 - In Control group, $33/52$
 - $OR = (41/44)/(33/52) = 1.47$
- Chi-squared statistic = 1.53, associated p-value = 0.22
- No evidence that Hodgkin's is associated with Tonsillectomy

(In)dependence: Johnson failed to account for pairing

Johnson et al.

	Tonsillectomy	No Tonsillectomy
Hodgkin's (n = 85)	41	44
Control (n = 85)	33	52

- This analysis IGNORED pairing (siblings and controls were *matched*)

	Sibling Tonsillectomy	Sibling No Tonsillectomy
Hodgkin's Tonsillectomy	26	15
Hodgkin's No Tonsillectomy	7	37

- Correct contingency table shows pairings (treats the unit of analysis as a pair)

(In)dependence: McNemar's Test

Johnson et al.

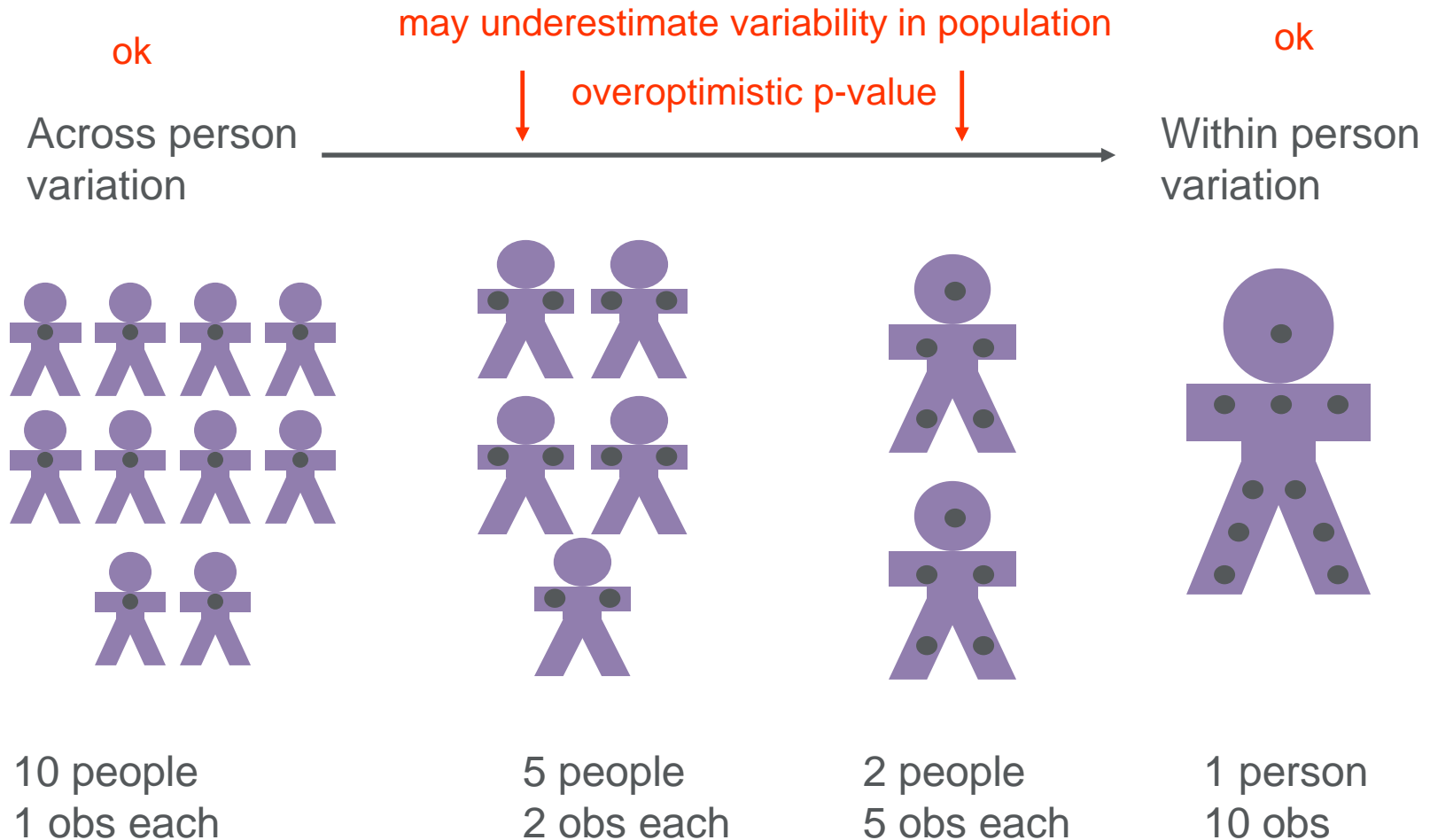
	Sibling Tonsillectomy	Sibling No Tonsillectomy
Hodgkin's Tonsillectomy	26	15
Hodgkin's No Tonsillectomy	7	37

- Chi-squared test WRONG choice
- Compare discordant pairs (McNemar's Test):
- Proportion of pairs in which sibling had tonsillectomy but Hodgkin's did not
 $7/85 = 8\%$
- Proportion of pairs in which sibling did not have tonsillectomy but Hodgkin's did
 $15/85 = 17\%$
- P-value 0.09
- Less doubt about results of Vianna et al.

(In)dependence: Think about types of variation

Across & Within Person Variation

If assume observations are independent ...



(In)dependence: Different Statistical Approaches

What you might use for independent data	What you might use for paired/dependent data
Chi-squared test	McNemar's test
Two-sample t-test	Paired t-test
Wilcoxon rank-sum test	Wilcoxon signed rank-sum test
Generalized Linear Model	Generalized Linear Mixed Model

NOTE: This is not a recipe for what to do if your data contains dependence, but rather an illustration of what MIGHT be suitable.

Minor Detail #3:

Assuming the mean is a
good measure of central
tendency

Defaulting to the Mean: Mean vs Median Example

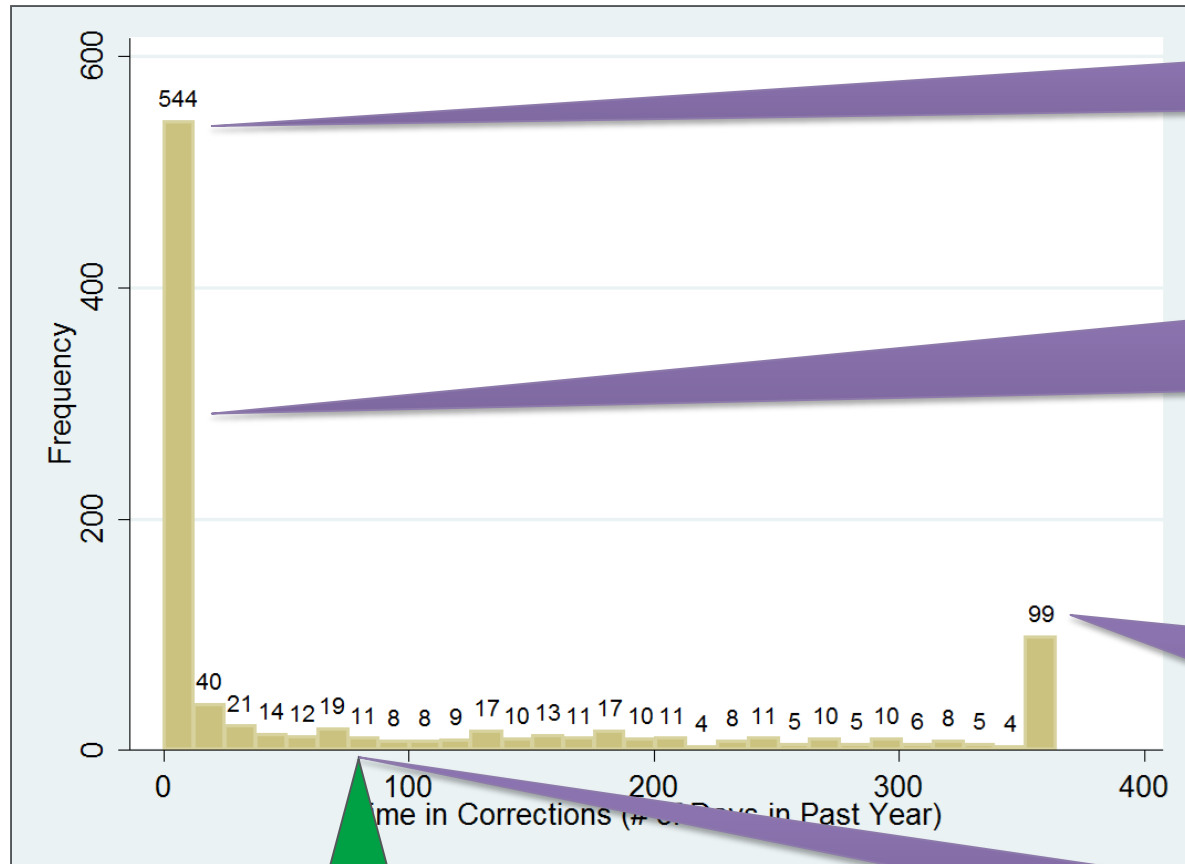
Examining time incarcerated in the past year

- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Looking at re-incarceration
- Goal is to summarize time incarcerated in the past year
 - Mean time incarcerated = 84 days
 - Median time incarcerated = 0 days

These are really
different
estimates -
what's going on?

Defaulting to the Mean: Mean vs Median Example

Look at the data



Over 50% of participants have no time incarcerated

Median is "middle" observation. N = 1000, 544 0's, so Median = 0 days

Some participants have very large values (365 days)

Mean is 'balance point' of distribution
84 days

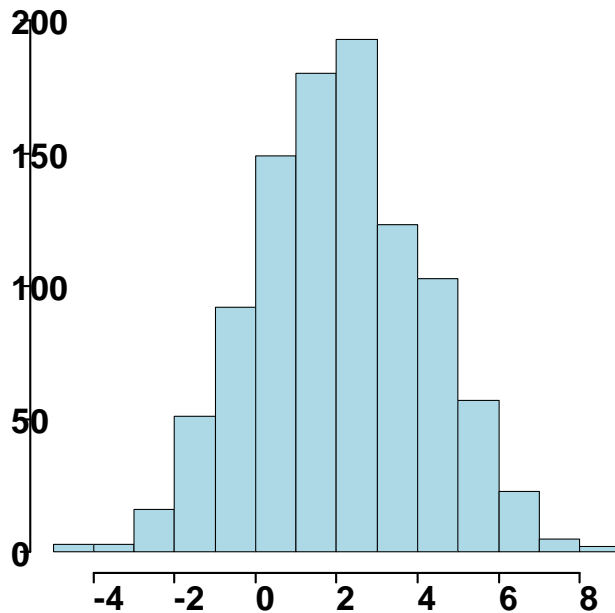
Defaulting to the Mean: Mean vs Median Example

What should you report when data are skewed?

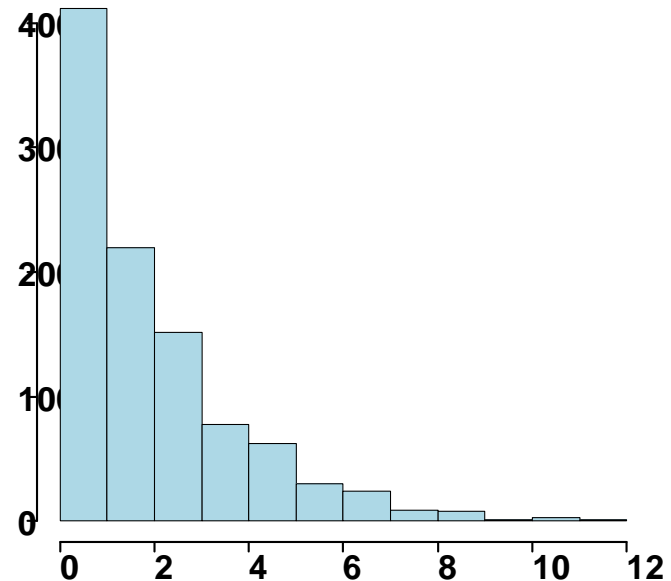
- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Looking at re-incarceration
- Goal is to summarize time incarcerated in the past year
 - Mean time incarcerated = 84 days
 - Median time incarcerated = 0 days
- What should we report?
 - People expect to see the mean (and the associated standard deviation)
 - I recommend also reporting the median, range, Q1, and Q3
- In this case, it may be better to separately
 - Report the fraction of participants who were never re-incarcerated
 - Report mean/median etc. among the 456 who we re-incarcerated

Defaulting to the Mean: Picture Your Data!

What do you think of when you hear “The mean value was 2.0”?



What we tend to think
Mean = 2
Median = 2



What might be true
Mean = 2.0
Median = 1.4

Defaulting to the Mean: SD vs SE

Averages are less variable than individual observations

- Standard deviation (SD) describes the *variability in a population*
- Standard error (SE) describes the *variability of an estimate from a sample*
- American women are on average 5'4", with a standard deviation of about 3"
 - Height is normally distributed, so approx 95% of women +/- 2 SD
 - 95% confidence interval for next woman to walk through the door

Describes variability in the population of American women

(4'10" – 5'10")

- Average height in a sample of 35 American women
 - Average is likely to be around 5'4"; with a standard error of $3/\sqrt{35} = 0.5$
 - 95% confidence interval for AVERAGE height of next 35 women through door

Describes variability in the mean of the sample of 35

(5'3" – 5'5")

Defaulting to the Mean: Recommendations

- The mean is not robust to outliers
- For skewed distributions, or distributions with outliers, the mean may be misleading
- In a manuscript, don't blindly report mean.
- Why use the mean at all?
 - Mathematically convenient
 - Nice statistical properties
- Standard deviation describes variability in a population, and standard error describes variability in an estimate from a sample

Minor Detail #4:

Using Excel for data
capture, cleaning, or
analysis

Using Excel: Potential problems for research

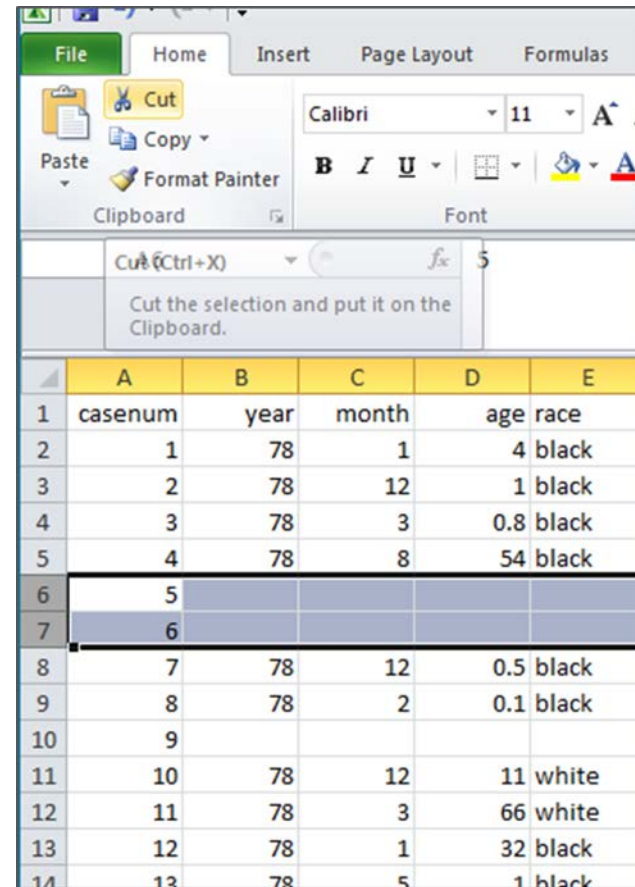
- Excel is available and accessible
- It's not uncommon for use with research data
 - Data capture
 - Data cleaning
 - Data analysis
 - Generating figures
- It's critical that we conduct rigorous and reproducible research
 - Excel not always optimal
- When is it okay to use Excel, and when is it not recommended?

Using Excel: Issues with data capture

Problems with entering data directly in to an Excel spreadsheet

Problems with entering data directly in to an Excel spreadsheet:

1. One-off/misalignment errors, especially for wide spreadsheets
2. Easy to unknowingly move or delete data
3. No explicit version control, trace-back, record or date stamp.
4. Standardization (e.g., Black vs black, blank vs “missing”)



The screenshot shows the Microsoft Excel interface. The ribbon includes File, Home, Insert, Page Layout, and Formulas. The Home ribbon is active, showing the Clipboard group with Cut, Copy, and Paste options, and the Font group with font face (Calibri), size (11), and bold/italic/underline options. A context menu is open over cell B6, showing 'Cut (Ctrl+X)' and a tooltip that reads 'Cut the selection and put it on the Clipboard.' The spreadsheet data is as follows:

	A	B	C	D	E
1	casenum	year	month	age	race
2	1	78	1	4	black
3	2	78	12	1	black
4	3	78	3	0.8	black
5	4	78	8	54	black
6	5				
7	6				
8	7	78	12	0.5	black
9	8	78	2	0.1	black
10	9				
11	10	78	12	11	white
12	11	78	3	66	white
13	12	78	1	32	black
14	13	78	5	1	black

Using Excel: Issues with data coding

Excel isn't designed for data that's coded, and has features that can lead to poor formatting for analysis

213	320285557	1	0		
214	320368722	1	0	1	Y, double adenoma
215	320368722	2	0	1	Y, double adenoma
216	320440974	1	0		
217	320468391	1	0		
218	320469342	1	0		
219	320581700	1	0		
220	321248969	1	0	0	N
221	321248969	2	0	0	N
222	321302497	1	1		
223	321304647	1	0		
224	321346362	1	0		
225	322163471	1	0		
226	323305945	1	0	1	Y, double adenoma
227	323305945	2	0	1	Y, double adenoma
228	323483774	1	0		
229	324106248	1	0		
230	324421569	1	0		
231	324480106	1	0	1	Y, single adenoma
232	324480106	2	0	1	Y, single adenoma

Yellow highlighting means something. But it is very hard to translate that in to a variable for analysis.

This column seems to contain two variables:

1. Adenoma present (Y/N)
2. Type of adenoma (single/double)

What's the difference between N and blank?

Are blanks missing values, unknown, or not applicable? All have different implications for analysis.

Mixing variable codes (Y/N) with plain text. It is very hard to tell a computer what to do with this, especially when you mistype adenoma.

Using Excel: Issues with data formatting

Gene names converted to dates or floating point numbers

Ziemann et al. *Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT

Open Access

Gene name errors are widespread in the scientific literature



Mark Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to ‘2-Sep’ and ‘1-Mar’, respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession ‘2310009E13’ to ‘2.31E+13’). Since

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and .xlsx suffixes) were converted to tabular separated files (tsv) with *ssconvert* (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Oryza sativa* and *Saccharomyces cerevisiae* [2]. The regex search used was similar to that described previously by Zeeberg and colleagues [1], with the added screen for dates in other formats (e.g. DD/MM/YY and MM-DD-YY). To expedite analysis of supplementary files from multi-disciplinary journals, we limited the articles screened to those that have the keyword ‘genome’ in the title or abstract (*Science*, *Nature* and *PLoS One*). Excel files (.xls and .xlsx) deposited

SEPT2 (Septin 2) converted to “2-Sept”
MARCH1 [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] converted to “1-Mar”

Using Excel: REDCap for capture, coding

Supports robust data capture and consistent data coding, formatting

- Research Electronic Data Capture
- Secure web application
- <http://project-redcap.org>
- Features:
 - Rapid set-up
 - Web-based data collection
 - Data validation
 - Export to statistical programs
 - Supports HIPAA compliance

The screenshot shows the REDCap website homepage. At the top left is the REDCap logo with the tagline "Research Electronic Data Capture". To the right, there is a "Consortium Wiki (Login Required)" link and a section for "Upcoming Events" listing a "Weekly All-Hands Consortium Meeting - Every Friday, 1-2PM Central". Below the header is a navigation menu with tabs for "Introduction", "Software", "Consortium Partners", "Become a Partner", "Video Resources", "Citing REDCap", and "Library". The main content area includes a paragraph about the consortium's 842 active institutional partners in 70 countries, a paragraph about the application's capabilities for building and managing online surveys, and a "Map of REDCap Consortium Partners" showing a world map with red location pins. To the right of the map is a "Recent publications using REDCap" section with several article titles and links. At the bottom of the page, it says "Powered by VANDERBILT" and "© 2013 Vanderbilt University".

Using Excel: REDCap for data capture, coding

REDCap vs Excel

Enrollment Assign record to a Data

Adding new Study ID 999

Event Name: **Baseline**

Study ID: 999

Enrollment Date:

Inclusion Criteria

Does the participant meet the definition of hypertensive (i.e., SBP/DBP >= 140/90)? Yes No

Is the participant 18 years of age or older? Yes No

Is the participant female of childbearing potential (i.e., pre-menopausal)? No, the participant is not of childbearing potential No, the participant is of childbearing potential Yes

Is the participant considered obese according to the study criteria (BMI at least 30 kg/m²)? Yes No

Does the participant agree to comply with all protocol-required study procedures? Yes No

Did the participant sign the study's informed consent document? Yes No

Does the participant have any pre-existing condition that, in the investigator's opinion, would preclude participation in the study? Yes No

Form Status

Complete? Incomplete Complete

Save Record
Save and Continue

File Home Insert Page Layout Formulas

Cut Copy Paste Format Painter Clipboard Font

Calibri 11

Cut (Ctrl+X) 5

Cut the selection and put it on the Clipboard.

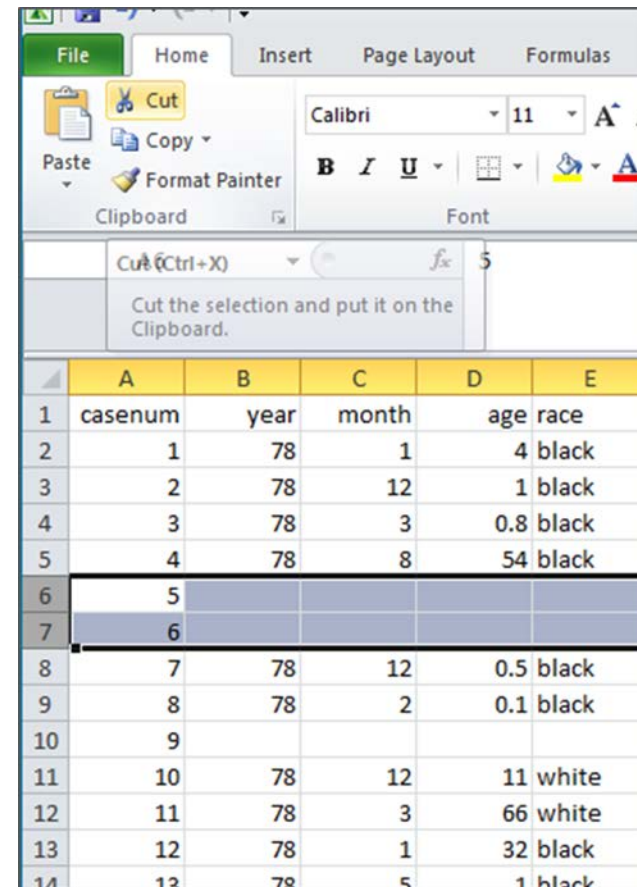
	A	B	C	D	E
1	casenum	year	month	age	race
2	1	78	1	4	black
3	2	78	12	1	black
4	3	78	3	0.8	black
5	4	78	8	54	black
6	5				
7	6				
8	7	78	12	0.5	black
9	8	78	2	0.1	black
10	9				
11	10	78	12	11	white
12	11	78	3	66	white
13	12	78	1	32	black
14	13	78	5	1	black

Using Excel: Issues with data cleaning/analysis

Problems with entering data directly in to an Excel spreadsheet

Problems with cleaning or analyzing data in to an Excel spreadsheet:

1. Repeated point-and-click, copy and paste, search and replace
2. No record of each step that was taken, and in what sequence (unless you write them all down)
3. Not very reproducible if there is a change to the original raw data, or questions about the analysis

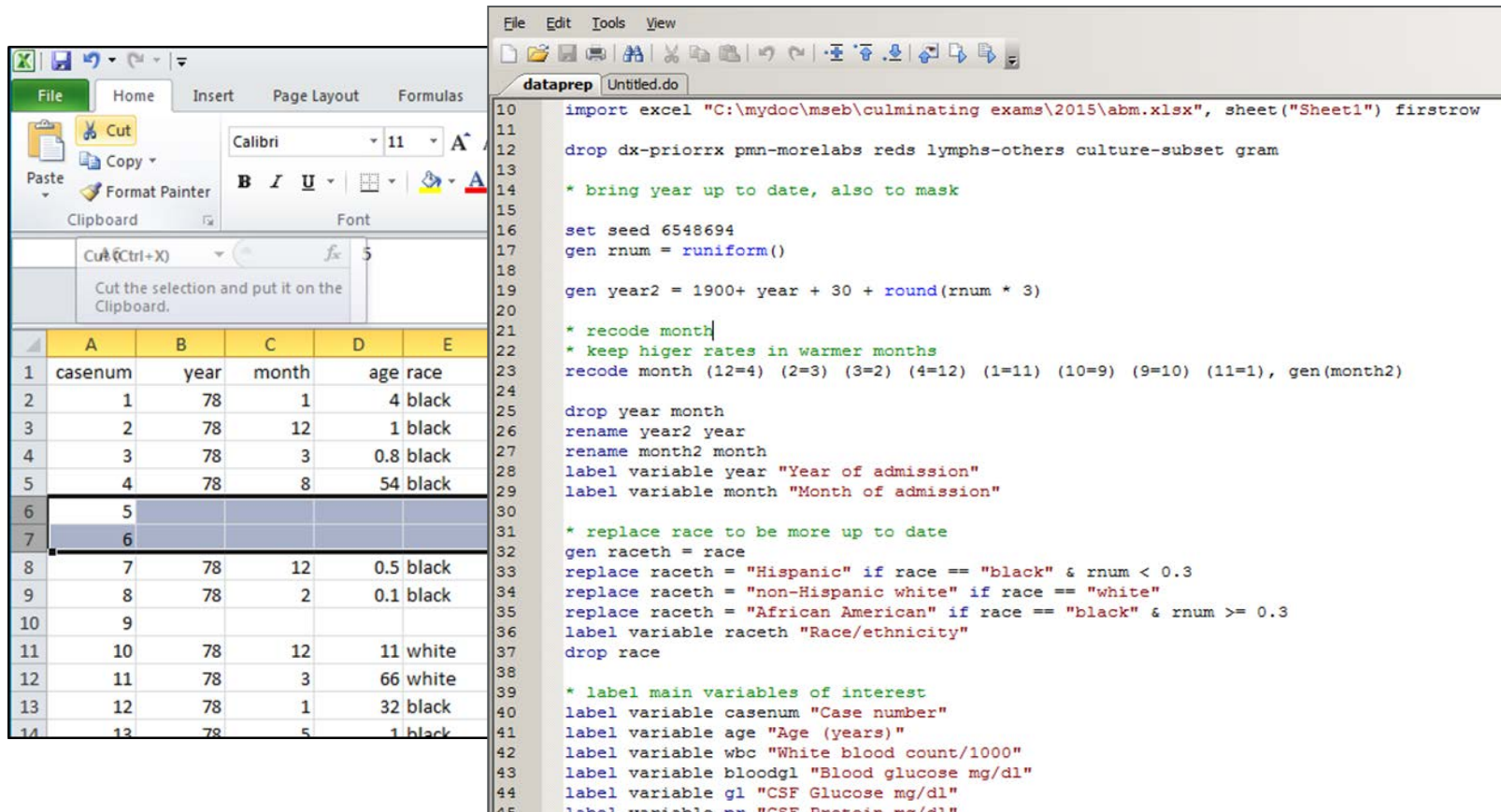


The screenshot shows the Microsoft Excel interface. The ribbon includes 'File', 'Home', 'Insert', 'Page Layout', and 'Formulas'. The 'Home' tab is active, showing the 'Clipboard' group with 'Cut', 'Copy', and 'Format Painter' buttons. A tooltip for the 'Cut' button is visible, stating 'Cut (Ctrl+X)' and 'Cut the selection and put it on the Clipboard.' The spreadsheet below has columns A through E and rows 1 through 14. The data is as follows:

	A	B	C	D	E
1	casenum	year	month	age	race
2	1	78	1	4	black
3	2	78	12	1	black
4	3	78	3	0.8	black
5	4	78	8	54	black
6	5				
7	6				
8	7	78	12	0.5	black
9	8	78	2	0.1	black
10	9				
11	10	78	12	11	white
12	11	78	3	66	white
13	12	78	1	32	black
14	13	78	5	1	black

Using Excel: Issues with data analysis

Cleaning/analyzing data in Excel versus statistical program



The image displays two side-by-side windows. On the left is a Microsoft Excel spreadsheet with a table of data. On the right is a statistical software interface (likely Stata) showing a script of commands for data cleaning and analysis.

	A	B	C	D	E
1	casenum	year	month	age	race
2	1	78	1	4	black
3	2	78	12	1	black
4	3	78	3	0.8	black
5	4	78	8	54	black
6	5				
7	6				
8	7	78	12	0.5	black
9	8	78	2	0.1	black
10	9				
11	10	78	12	11	white
12	11	78	3	66	white
13	12	78	1	32	black
14	13	78	5	1	black

```
File Edit Tools View
dataprep Untitled.do
10 import excel "C:\mydoc\mseb\culminating exams\2015\abm.xlsx", sheet("Sheet1") firstrow
11
12 drop dx-priorrx pmn-morelabs reds lymphs-others culture-subset gram
13
14 * bring year up to date, also to mask
15
16 set seed 6548694
17 gen rnum = runiform()
18
19 gen year2 = 1900+ year + 30 + round(rnum * 3)
20
21 * recode month|
22 * keep higer rates in warmer months
23 recode month (12=4) (2=3) (3=2) (4=12) (1=11) (10=9) (9=10) (11=1), gen(month2)
24
25 drop year month
26 rename year2 year
27 rename month2 month
28 label variable year "Year of admission"
29 label variable month "Month of admission"
30
31 * replace race to be more up to date
32 gen raceth = race
33 replace raceth = "Hispanic" if race == "black" & rnum < 0.3
34 replace raceth = "non-Hispanic white" if race == "white"
35 replace raceth = "African American" if race == "black" & rnum >= 0.3
36 label variable raceth "Race/ethnicity"
37 drop race
38
39 * label main variables of interest
40 label variable casenum "Case number"
41 label variable age "Age (years)"
42 label variable wbc "White blood count/1000"
43 label variable bloodgl "Blood glucose mg/dl"
44 label variable gl "CSF Glucose mg/dl"
45 label variable pr "CSF Protein mg/dl"
```

Inefficient and potentially inaccurate to repeat cleaning/analysis in Excel. With scripted code, it's easy to re-run a data cleaning or analysis program.

Using Excel: Alternatives such as R Studio, Stata

Getting more user friendly, and much more robust than Excel

Model Terms by/if/in Weights Options Reporting

Dependent variable: Independent variables:

Independent variables: (do not report coefficients)

lifetime_linreg - Notepad

```

name: <unnamed>
log: P:\Products\Papers\Kid's Papers\Kid's dev sub\Analysis\log\lifetime_linreg.log
Log type: text
opened on: 24 Apr 2013, 13:09:07

.*sedative use
. qui wt_corr o_seddsmlf
0.22%
. svy, sub(if male==1 & race==1): logistic o_seddsmlf black hisp
(running logistic on estimation sample)

Survey: Logistic regression
Number of strata = 9
Population size = 1165
Subpop. no. of obs = 1165
Subpop. size = 1684.3324
Design df = 1156
F( 2, 1155) = 23.43
Prob > F = 0.0000

-----
o_seddsmlf | odds Ratio | Linearized Std. Err. | t | P>|t| | [95% Conf. Interval]
-----+-----
black | .0030322 | .0031084 | -5.66 | 0.000 | .0004057 | .0226608
hisp | .1025896 | .0535013 | -4.37 | 0.000 | .0368748 | .2854154
_cons | .1309232 | .0292753 | -9.09 | 0.000 | .0844272 | .2030256
-----
Note: 4 strata omitted because they contain no subpopulation members.

. svy, sub(if male==1 & race==1): logistic o_seddsmlf white hisp
(running logistic on estimation sample)
    
```

RStudio

Home RStudio IDE Shiny Training Projects About Blog

RStudio IDE

About Screenshots Download Documentation Support Development

Take control of your R code

RStudio is a free and open source integrated development environment for R. You can run it on your desktop (Windows, Mac, or Linux) or even over the web using RStudio Server.

Download RStudio
for Windows, Mac, or Linux

Screencast

R Studio is a way to use R statistical software (free, open source) in a user friendly environment with more point-and-click capability.

Stata has menus that you can use to point and click, but it will generate the statistical code file for you and keep a log of all your work!

Using Excel: poster child for why not to

A disastrous story in why not to use Excel

The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291
© Institute of Mathematical Statistics, 2009

DERIVING CHEMOSENSITIVITY FROM FORENSIC BIOINFORMATICS AND REPLICATED RESEARCH IN HIGH-THROUGHPUT

BY KEITH A. BAGGERLY¹ AND KEVIN R. COOPER

University of Texas

High-throughput biological assays such as microarray
detailed questions about how diseases operate, and promise
alize therapy. Data processing, however, is often not descri
to allow for exact reproduction of the results, leading to exer
bioinformatics” where aspects of raw data and reported resu
fer what methods must have been employed. Unfortunately,
tion can shift from an inconvenience to an active danger wh
just methods but erro
reporting to use micro
cell lines to predict p
being allocated to tre
show in five case stu
that may be putting p
common errors are si
experience that the m
are taking to avoid such errors in our own investigations.

Misconduct in science

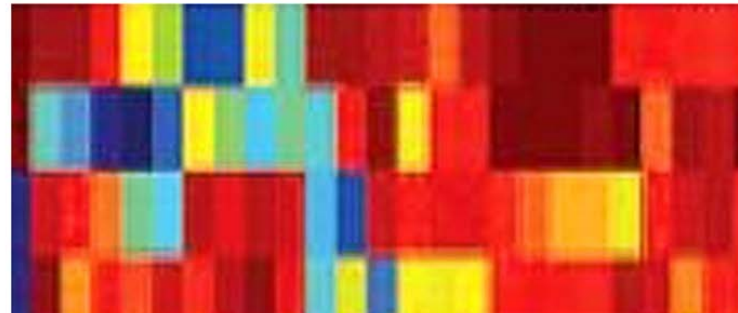
An array of errors

Investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing

Sep 10th 2011 | From the print edition

Like 962

Tweet 217



“The most simple problems are common.” When using Excel, it is especially easy to make off-by-one errors (e.g. accidentally deleting a cell in one column), or mixing up group labels (e.g. swapping sensitive/resistant).

that they had developed a similar technique which used gene expression in laboratory cultures of cancer cells, known as cell lines, to predict which chemotherapy would be most effective for an individual patient suffering from lung, breast or ovarian cancer.

At the time, this work looked like a tremendous advance for personalised medicine—the idea that understanding the molecular specifics of an individual's illness will lead to a tailored treatment. The papers drew adulation from other workers in the field, and many newspapers, including this one ([see article](#)), wrote about them. The team then started to organise a set of clinical trials of personalised treatments for lung and breast cancer. Unbeknown to most people in the field, however, within a few weeks of the publication of the

Using Excel: Recommendations

- *Try to avoid* capturing, manipulating, and analyzing your data in Excel
- Be careful when ‘parking’ your data in Excel
 - Data is often passed around in .csv format, which Excel easily reads
 - Excel isn’t bad per se for viewing .csv data
- Data cleaning, reshaping can eat up a lot of analysis time, sometimes more than the analysis itself, so the investment of time up front is worth it
- In the conduct of rigorous, reproducible research, Excel can be a weak link

Minor Detail #5:

Big statistics for little data:
Right-sizing the statistical
approach to the sample size

Right-sizing the statistics: big ideas, little data

- It's tempting to come up with sophisticated models, but need to think about whether or not you have enough data to explore them.
- Especially relevant when proposing your main hypothesis for a study
- For example, proposing mediation models (see upcoming lecture) for a sample of $n = 100$
 - Most statisticians will raise their eyebrows
 - Too small unless you're detecting pretty big effects
- Sometimes even simple comparisons require a lot of data.
 - E.g. Comparing prevalence between two groups
 - Expect prevalence in one group is 3%, and other group is 8%
 - Still need more than 300 per group to detect this difference as significant with 80% power

Right sizing the statistics: A (Very) General Rule

If you're comparing a binary (yes/no) outcome, you need at least 10 observations of each type (yes/no) per "degree of freedom" to have a reasonable chance at *estimating* those differences.

Can think of a "degree of freedom" as a variable you will put in your logistic regression model.

This does not guarantee *any* sort of power!

Independent Variable	Degrees of Freedom	Minimum Sample Size*
Sex (Man, Woman)	1	20
Age (continuous)	1	20
Age (< 20, 20-40, 40-60, 60+)	3	60
Race/Ethnicity (non-Hispanic white, African American, Hispanic, Asian, Other)	4	80
Sex + Race/Ethnicity	5	100

* Assumes an even split, so you probably need a lot more.

Right-sizing the statistics: small samples

Sometimes see 'regular' statistical approaches applied to small data sets

- E.g. Two sample t-test comparing groups of 15 each

There are alternatives:

- Non-parametric approaches
 - Makes fewer distributional assumptions about the data
 - E.g. Fisher's exact test instead of a chi-squared test
- Exact approaches
 - Same models, but estimated differently
 - E.g. "Exact" logistic regression vs (maximum likelihood) logistic regression
- Bootstrapping
 - Resampling your data to obtain better standard errors
 - Doesn't always solve the small sample problem

Right-sizing the statistics: Effect sizes

Effect sizes are *relative*

- Power/sample size considerations often calculated in terms of ‘effect size’

Power	Total N	Effect size
0.8	788	0.2 (“small”)
0.8	128	0.5 (“medium”)
0.8	52	0.8 (“large”)
0.8	12	2.0

Two independent samples comparison of means with $\alpha = 0.05$.

- Effect ‘size’ is relative to the standard deviation of the outcome
- If SD of outcome is 10 units
 - Can detect a “small” difference of 2 units ($= 0.2 * 10$) with $n = 788$
 - Can detect a “medium” difference of 5 units ($= 0.5 * 10$) with $n = 128$
 - Can detect a “large” difference of 8 units ($= 0.8 * 10$) with $n = 52$
 - Can detect a “huge” difference of 20 units ($= 2.0 * 10$) with $n = 12$

Your feedback is important to us! (And helps us plan future lectures).

Complete the evaluation survey to be entered in to a drawing to win 2 free hours of biostatistics consultation.

Statistically Speaking: Upcoming Lectures

We hope to see you again!

Monday, October 9 **The Impact of Other Factors: Confounding, Mediation, and Effect Modification**
Amy Yang, MS, Sr. Statistical Analyst, Division of Biostatistics,
Department of Preventive Medicine

Monday, October 16 **Using REDCap for Data Capture in Clinical Studies: Database Management on a Budget**
Jody D. Ciolino, PhD, Assistant Professor, Division of Biostatistics,
Department of Preventive Medicine

Monday, October 30 **Using R for Statistical Graphics: The Do's and Don'ts of Data Visualization**
David Aaby, MS, Sr. Statistical Analyst, Division of Biostatistics,
Department of Preventive Medicine

Wednesday, November 1 **Time-to-Event Analysis: A 'Survival' Guide**
Lauren C. Balmert, PhD, Assistant Professor, Division of Biostatistics,
Department of Preventive Medicine

All lectures will be held from noon to 1 pm in Baldwin Auditorium, Robert H. Lurie Medical Research Center, 303 E. Superior St.

BCC: Biostatistics Collaboration Center

Contact Us

- Request an Appointment
 - <http://www.feinberg.northwestern.edu/sites/bcc/contact-us/request-form.html>
- General Inquiries
 - bcc@northwestern.edu
 - 312.503.2288
- Visit Our Website
 - <http://www.feinberg.northwestern.edu/sites/bcc/index.html>

Biostatistics Collaboration Center |680 N. Lake Shore Drive, Suite 1400 |Chicago, IL 60611