

Four Lesson in Statistics, and Some Statistics Along the Way

STATISTICALLY SPEAKING ...

October 2016

Leah J. Welty, PhD

Director, Biostatistics Collaboration Center

Associate Professor

Department of Preventive Medicine, Division of Biostatistics

Department of Psychiatry and Behavioral Sciences

BCC: Biostatistics Collaboration Center

Who We Are



Leah J. Welty, PhD
Assoc. Professor
BCC Director



Joan S. Chmiel, PhD
Professor



Jody D. Ciolino, PhD
Asst. Professor



Kwang-Youn A. Kim, PhD
Asst. Professor



Masha Kocherginsky, PhD
Assoc. Professor



Mary J. Kwasny, ScD
Assoc. Professor



Julia Lee, PhD, MPH
Assoc. Professor



Alfred W. Rademaker, PhD
Professor



Hannah L. Palac, MS
Senior Stat. Analyst



Gerald W. Rouleau, MS
Stat. Analyst



Amy Yang, MS
Senior Stat. Analyst

- Not Pictured:
1. David A. Aaby, MS
Senior Stat. Analyst
 2. Tameka L. Brannon
Financial | Research
Administrator

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

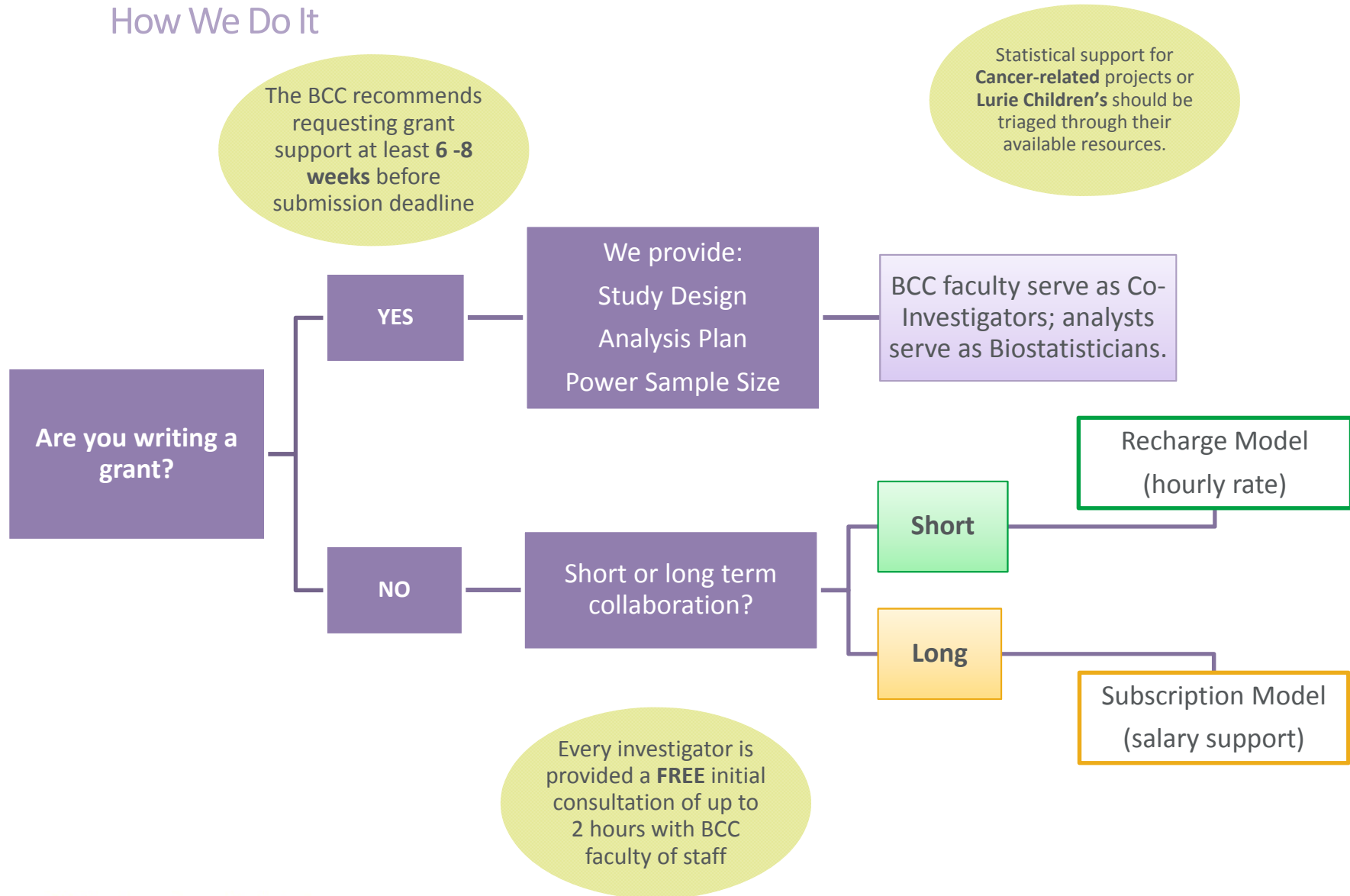
BCC: Biostatistics Collaboration Center

What We Do

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

BCC: Biostatistics Collaboration Center

How We Do It





BCC: Biostatistics Collaboration Center

How can you 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



Four Lessons in Statistics: Outline

4. A (good) picture is worth 1,000 words.
3. Not all observations are independent.
2. Are 50% of us really above average?
1. What the ** is a p-value?

Lesson #4

“A (good) picture is worth
1,000 words.”



Good Pictures: An example of why they're needed

Correlation Defined

Correlation often denoted as r

Measures strength of the *linear* association between two continuous variables

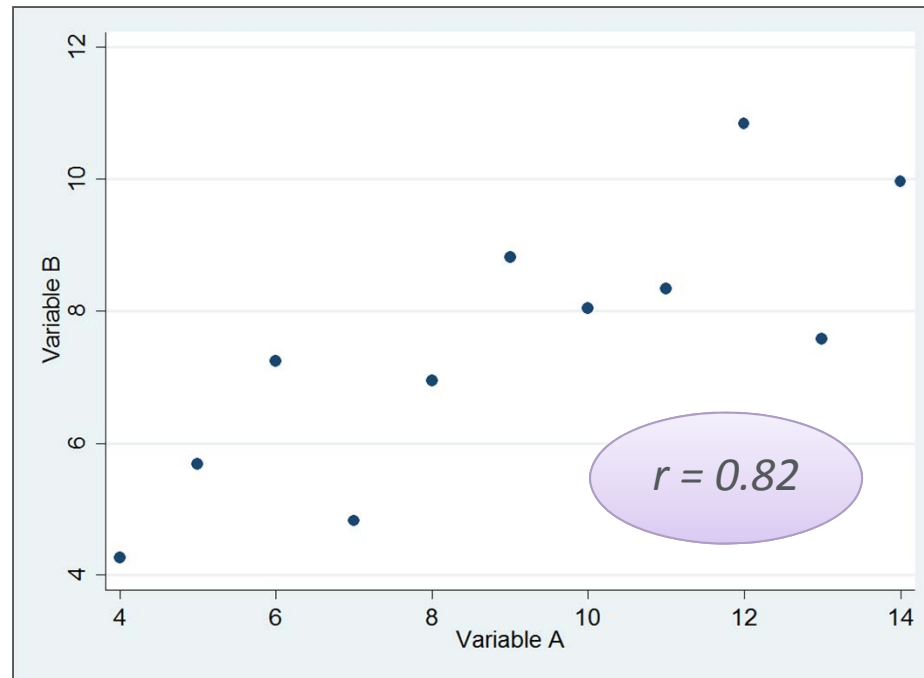
$$-1 \leq r \leq 1$$

- $r = -1$ strong negative linear association
- $r = 1$ strong positive linear association
- $r = 0$ no linear association

Good Pictures: An example of why they're needed

Correlation Example

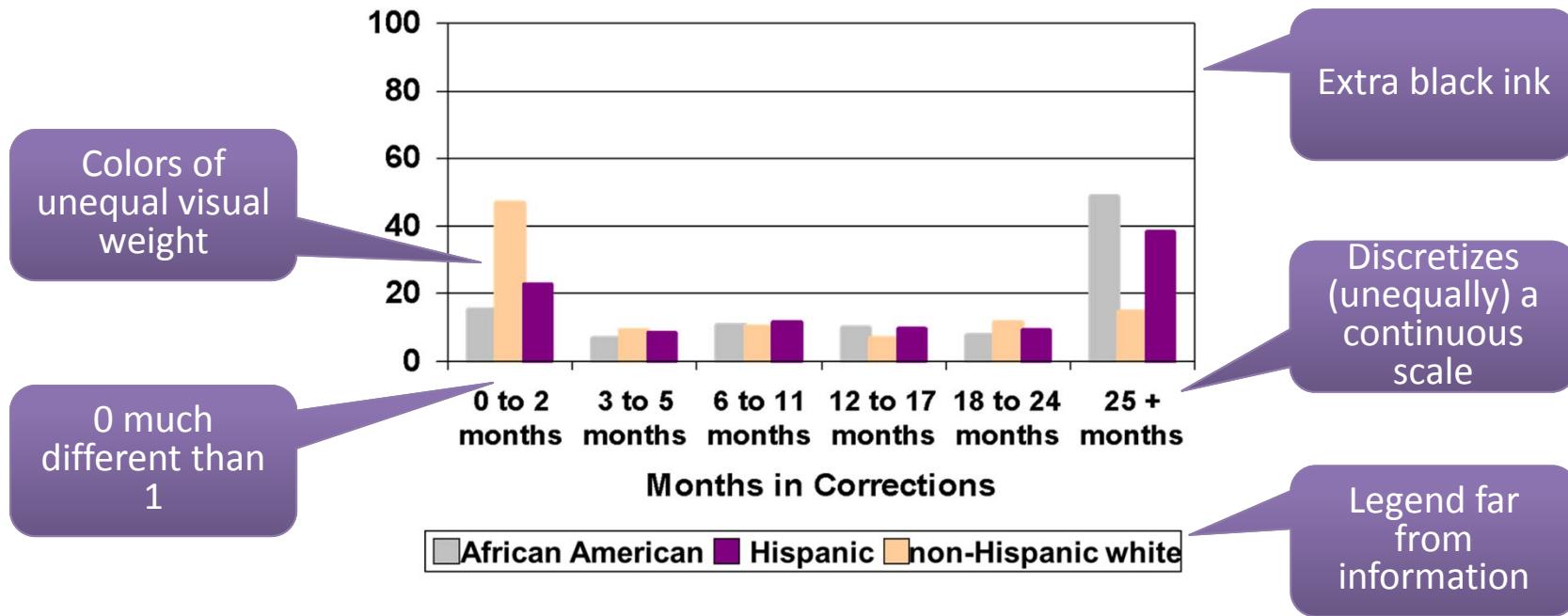
- The correlation between two variables of interest, A and B, is 0.82.
- Is there a strong positive linear association between A and B?



Good Pictures: Excel can lead you far astray

Bad pictures can hurt your brain

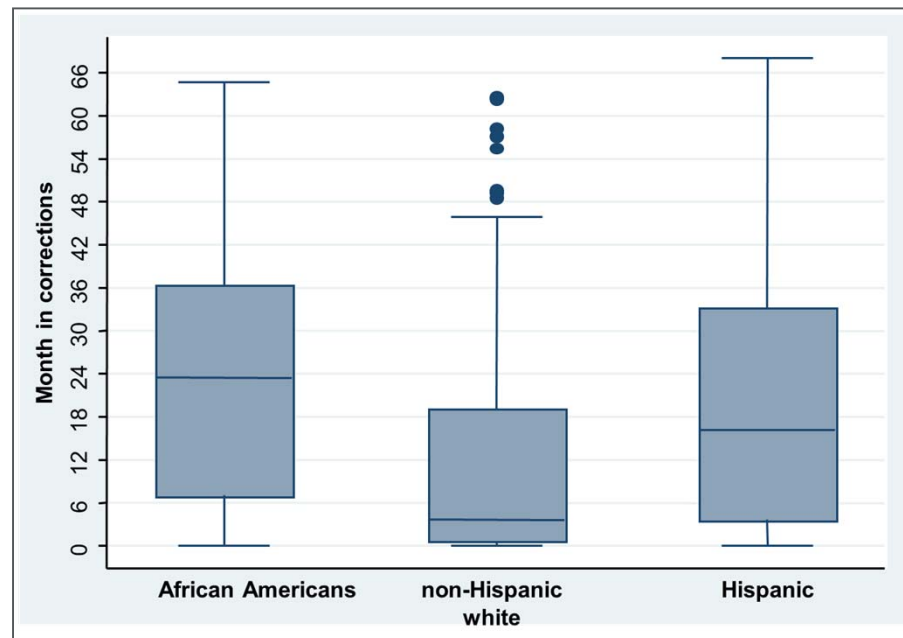
- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Are there racial/ethnic differences in
 - Being re-incarcerated
 - Length of time incarcerated



Good Pictures: Side-by-side boxplots

Simple but really useful

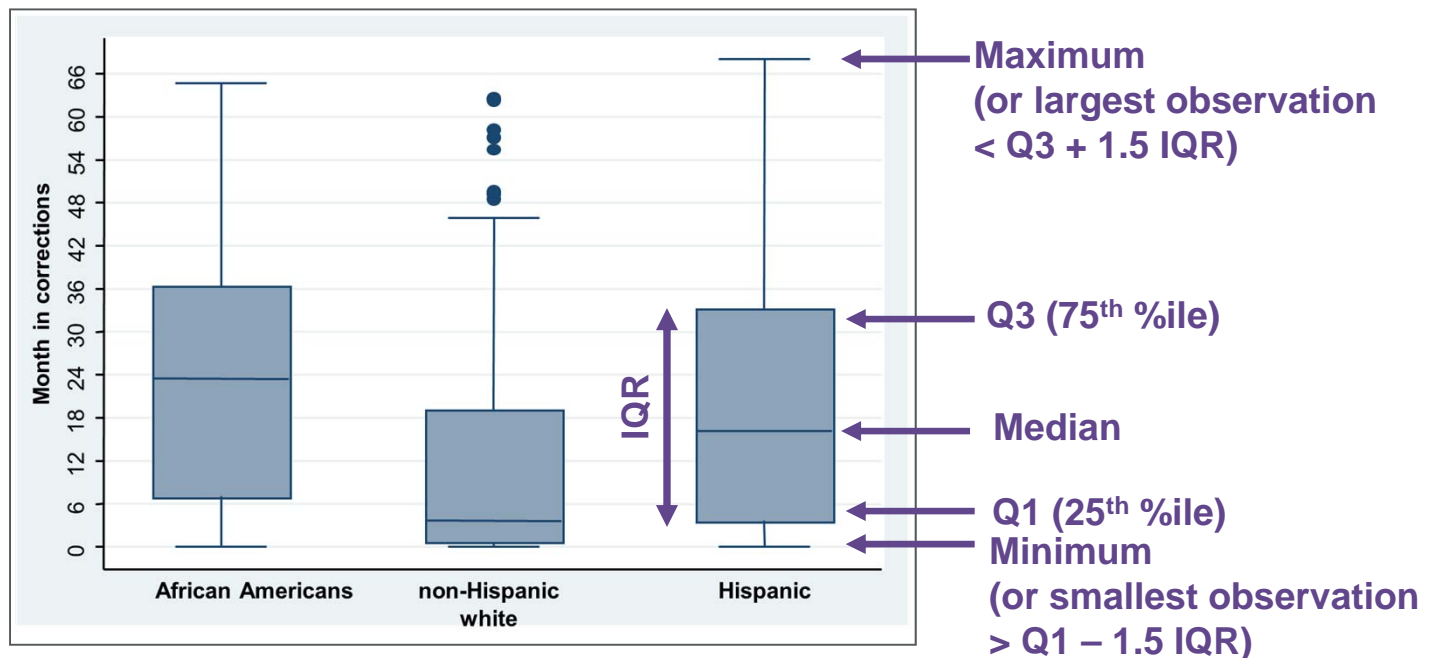
- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Are there racial/ethnic differences in
 - Being re-incarcerated
 - Length of time incarcerated



Good Pictures: Side-by-side boxplots

Simple but really useful

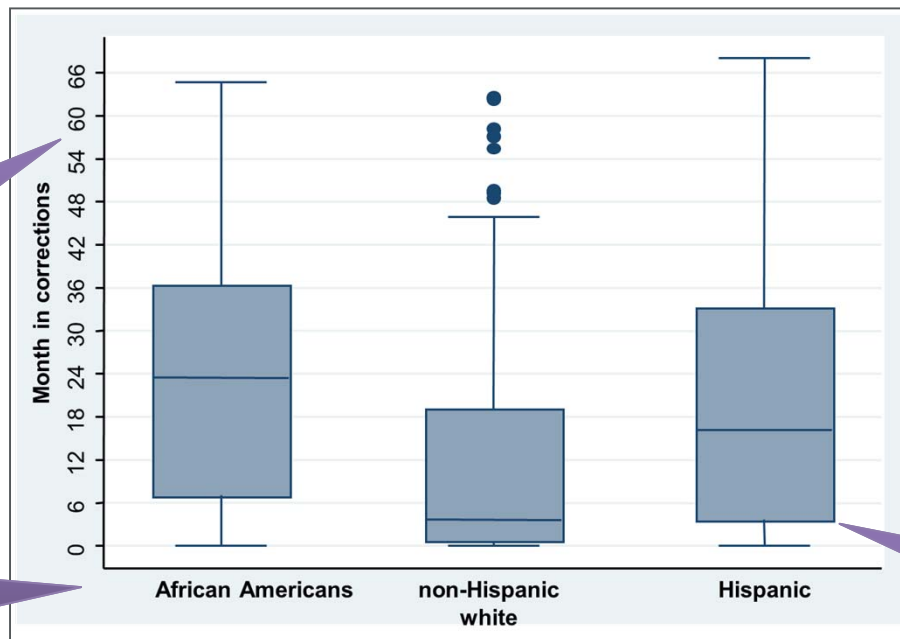
- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Are there racial/ethnic differences in
 - Being re-incarcerated
 - Length of time incarcerated



Good Pictures: Side-by-side boxplots

Simple but really useful

- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Are there racial/ethnic differences in
 - Being re-incarcerated
 - Length of time incarcerated



Months now continuous scale

Direct comparison of groups

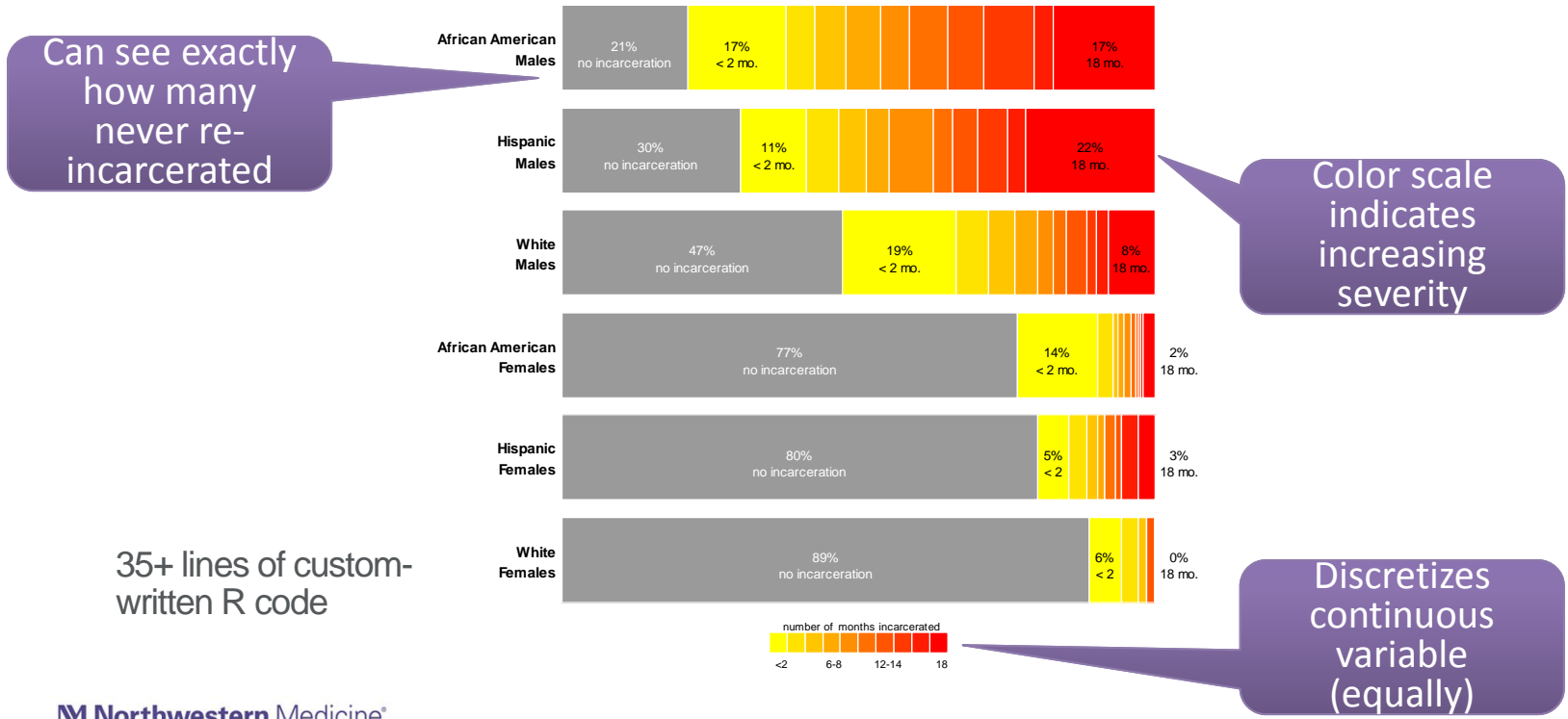
At least 75% re-incarcerated

Stata command: `graph box incarceration, over(race)`

Good Pictures: Custom graphics with R

Good pictures tell a story and you can get creative

- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Are there racial/ethnic differences in
 - Being re-incarcerated
 - Length of time incarcerated (*slightly different data*)





Good Pictures: Graphics “Rules”

Other points to keep in mind

- Maximum information; minimum ink (see work by Edward Tufte)
 - Tower and antenna plots waste a lot of ink
 - Color should be informative, thoughtful, and not gratuitous
- Graphics should have no more dimensions than exist in your data
 - No 3-d histograms
 - Only 3-d if you are plotting a surface
- Labels should be informative but not distracting
 - Graphics should stand on their own



Good Pictures: Worth 1,000 Statistics

Take home messages

- A good picture of your data
 - May help identify appropriate statistical methods
 - May help identify errors or irregularities
- A really good picture of your data
 - Can tell your story for you
 - Doesn't have to be complicated

Lesson #3

“Not all observations are independent.”



(In)dependence: 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

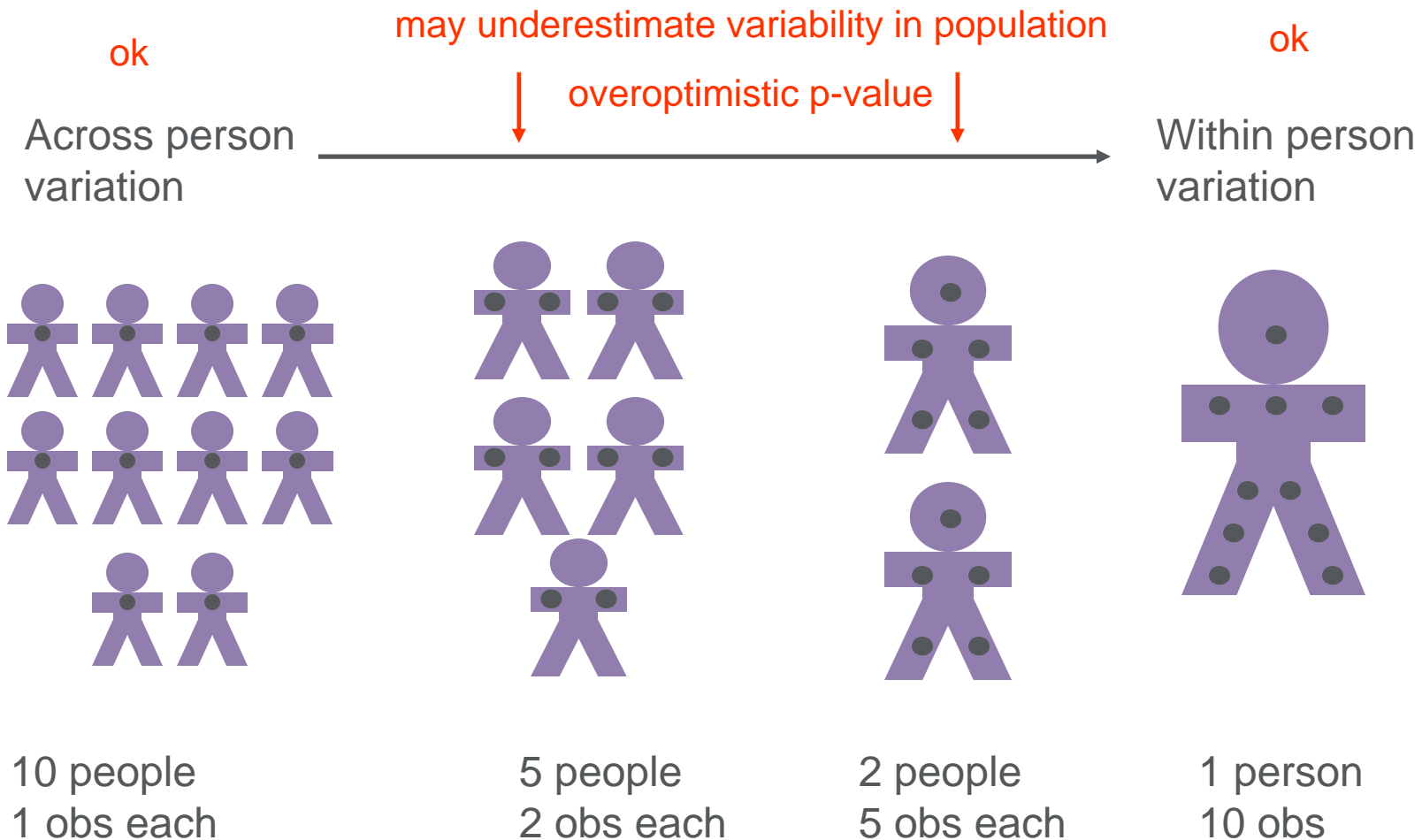
- Chi-squared test WRONG choice
- Compare discordant pairs (McNemar's Test):
- Proportion of pairs in which sibling had tonsillectomy but Hodgkin's did not
 $7/85 = 8\%$
- Proportion of pairs in which sibling did not have tonsillectomy but Hodgkin's did
 $15/85 = 17\%$
- P-value 0.09
- Less doubt about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



(In)dependence: Think about types of variation

Across & Within Person Variation

If assume observations are independent ...





(In)dependence: Recommendations

- Many common statistical methods assume observations are independent
- 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
- Paired data can be powerful and efficient, but you have to do your analysis right

Lesson #2

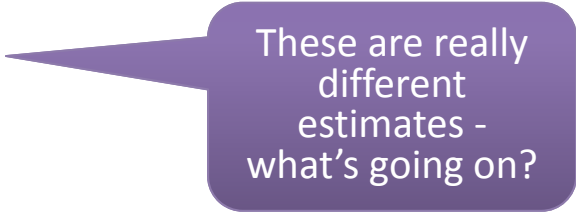
“Are 50% of us really above average?”



Above Average: Mean vs Median Example

Again examining time incarcerated in the past year

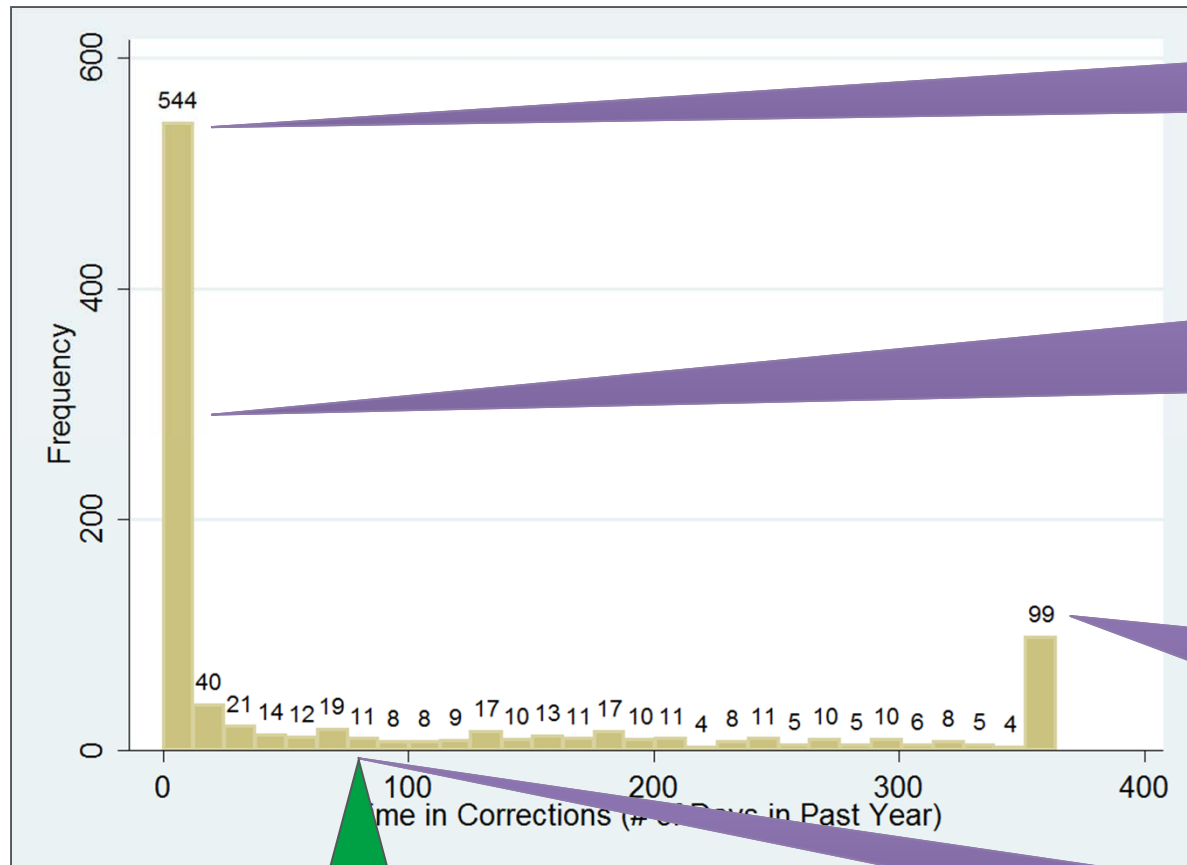
- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Again looking at re-incarceration
- Goal is to summarize time incarcerated in the past year
 - Mean time incarcerated = 84 days
 - Median time incarcerated = 0 days



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

Above Average: Mean vs Median Example

Look at the data! (Reprise to Lesson #4)



Over 50% of participants have no time in corrections

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



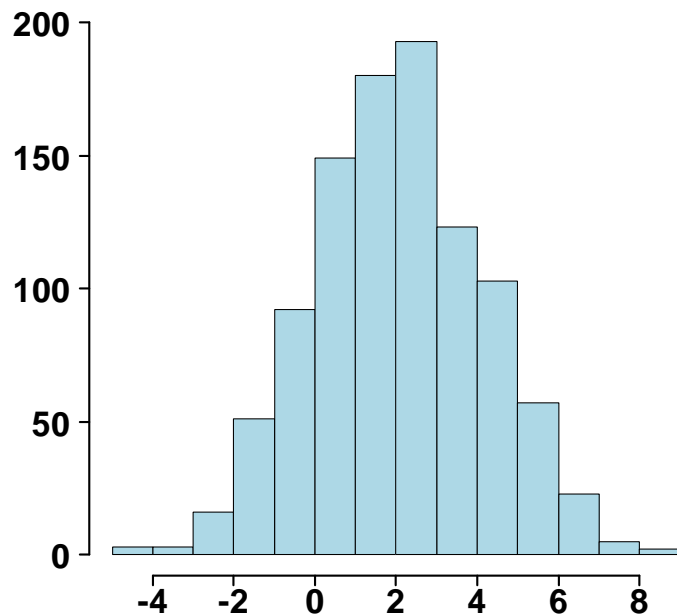
Above Average: Mean vs Median Example

What should you report when data are skewed?

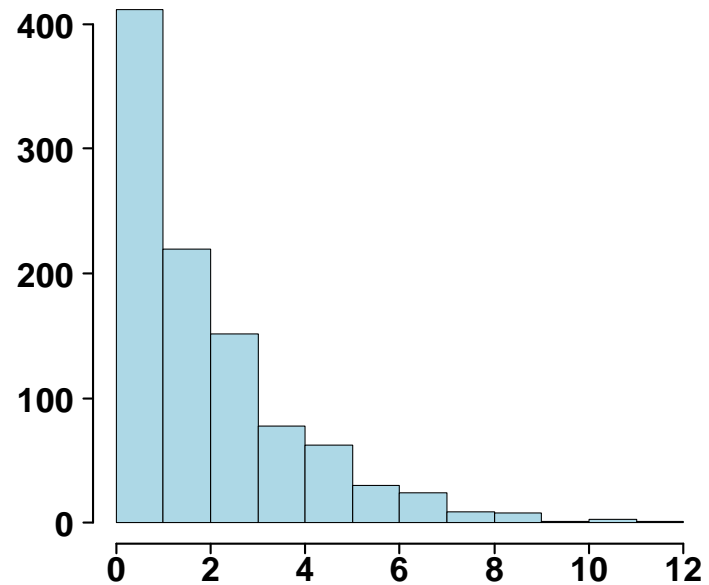
- Longitudinal study of juvenile delinquents (*Northwestern Juvenile Project*)
- Again looking at re-incarceration
- Goal is to summarize time incarcerated in the past year
 - Mean time incarcerated = 84 days
 - Median time incarcerated = 0 days
- What should we report?
 - People expect to see the mean (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

Above Average: 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

Above Average: Standard Deviation vs Standard Error

Averages are less variable than individual observations

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

Describes variability in the population of American women

(4'10" – 5'10")

Describes variability in the mean of the sample of 35

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

(5'3" – 5'5")



Above Average: Reminders and 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
- You are above average if you understand the important differences between the median and the mean, standard deviation and standard error

Lesson #1

“What the **** is a p-value?”



P-Values: An Analogy

Cheating at Poker

- Dr. X and I are playing poker
- Dr. X is beating me
- Dr. X's two most recent hands were a flush and a straight
- Is Dr. X cheating?



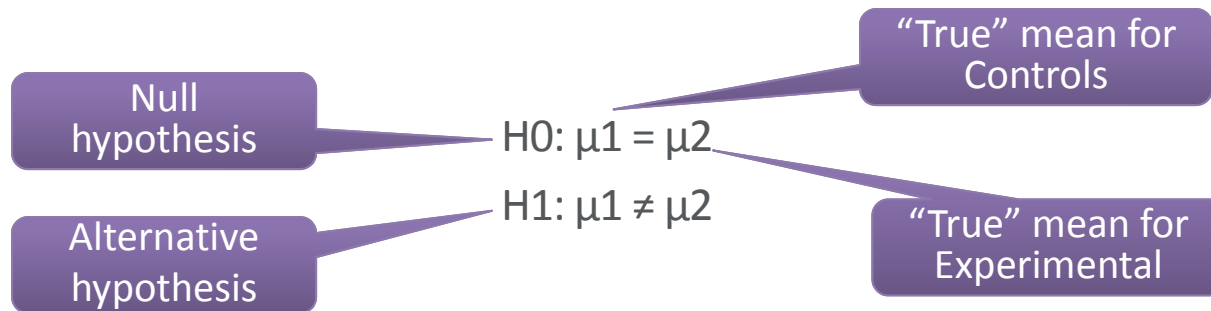
P-Values: Poker and Hypothesis Testing

A statistical approach to detect cheating

- Suppose Dr. X is playing fairly (opposite of what I suspect).
 - Called the *null hypothesis*, or H_0
- Observe the data.
 - Dr. X's next hand is 2 pair.
- What is the probability of Dr. X having a hand that is 2 pair or better if Dr. X is playing fairly?
 - Called the *p-value* (approximately 0.08 for this example)
- If probability is “small”, conclude that supposition might not be right.
 - Reject the null hypothesis in favor of the alternative hypothesis, or H_1
 - Conclude that evidence may support Dr. X cheating.
 - Conclusion may be wrong (what if Dr. X is very lucky/skillful?)
- If the probability is not “small”, conclude do not have evidence to reject the null hypothesis
 - Not the same as ‘accepting’ the null hypothesis, or showing that the null hypothesis is true
 - Dr. X may really be cheating, we just didn't ‘detect’ it.

P-Values: A more traditional example

Suppose you have a treatment that you suspect may alter performance on a task. You compare the means of your control and experimental groups (say, 20 subjects per group). You use a simple independent means t-test and your result is significant ($t = 2.7$, $df = 18$, $p = 0.01$).



Reject H_0 in favor of H_1 .

From "The Null Ritual: What you always wanted to know about significance testing but were afraid to ask." Gigerenzer, G., Krauss, S., Vitouch, O. in The Sage Handbook of Quantitative Methodology of the Social Sciences (2004). David Kaplan, Editor.

P-Values: Definition

- The p-value is the probability of the observed data (or of more extreme data), given that the null hypothesis H_0 is true.

$$\text{p-value} = \Pr(\text{data} \mid H_0)$$

← “given”

- This doesn't tell us what we might like to know: $\Pr(H_0 \mid \text{data})$ or $\Pr(H_1 \mid \text{data})$.



P-Values: Definition

- If Dr. X is not cheating, we would expect Dr. X to get a hand this good or better less than 8% of the time.

There is no $\Pr(\text{Dr. X Cheating})$. Dr. X is either cheating or not cheating.

- If there were no difference in means between the two groups, we would expect to see a difference in group means this large – or larger – about 1% of the time.

There is no $\Pr(\text{group means different})$.



P-Values: A Significance Test

Suppose you have a treatment that you suspect may alter performance on a task. You compare the means of your control and experimental groups (say, 20 subjects per group). You use a simple independent means t-test and your result is significant ($t = 2.7$, $df = 18$, $p = 0.01$).

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Reject H_0 in favor of H_1 .

From “The Null Ritual: What you always wanted to know about significance testing but were afraid to ask.” Gigerenzer, G., Krauss, S., Vitouch, O. in The Sage Handbook of Quantitative Methodology of the Social Sciences (2004). David Kaplan, Editor.

P-Values: A Significance Test

Answer each of the following true or false, recall $p = 0.01$

Statement	T/F ?
You have disproved the null hypothesis (i.e. there is no difference between population means).	
You have found the probability of the null hypothesis being true.	
You have proved your alternative hypothesis (i.e. that there is a difference between the population means).	
You can deduce the probability of the alternative hypothesis being true.	
If you reject the null hypothesis, you know the probability that you are making the wrong decision.	
If the experiment were repeated thousands of times, you would obtain a significant result $\sim 99\%$ of the time.	

P-Values: A Significance Test

Answer key

Statement	T/F ?
You have disproved the null hypothesis (i.e. there is no difference between population means).	F
You have found the probability of the null hypothesis being true.	
You have proved your alternative hypothesis (i.e. that there is a difference between the population means).	F
You can deduce the probability of the alternative hypothesis being true.	
If you reject the null hypothesis, you know the probability that you are making the wrong decision.	
If the experiment were repeated thousands of times, you would obtain a significant result ~ 99% of the time.	

The p-value is a probability, not proof.

P-Values: A Significance Test

Answer key

Statement	T/F ?
You have disproved the null hypothesis (i.e. there is no difference between population means).	F
You have found the probability of the null hypothesis being true.	F
You have proved your alternative hypothesis (i.e. that there is a difference between the population means).	F
You can deduce the probability of the alternative hypothesis being true.	F
If you reject the null hypothesis, you know the probability that you are making the wrong decision.	
If the experiment were repeated thousands of times, you would obtain a significant result ~ 99% of the time.	

p-value = $p(\text{data} | H_0)$ NOT $p(H_0 | \text{data})$ or $p(H_1 | \text{data})$

P-Values: A Significance Test

Answer key

Statement	T/F ?
You have disproved the null hypothesis (i.e. there is no difference between population means).	F
You have found the probability of the null hypothesis being true.	F
You have proved your alternative hypothesis (i.e. that there is a difference between the population means).	F
You can deduce the probability of the alternative hypothesis being true.	F
If you reject the null hypothesis, you know the probability that you are making the wrong decision.	F
If the experiment were repeated thousands of times, you would obtain a significant result ~ 99% of the time.	

p-value = p(data | H0) NOT p(H0)

P-Values: A Significance Test

Answer key

Statement	T/F ?
You have disproved the null hypothesis (i.e. there is no difference between population means).	F
You have found the probability of the null hypothesis being true.	F
You have proved your alternative hypothesis (i.e. that there is a difference between the population means).	F
You can deduce the probability of the alternative hypothesis being true.	F
If you reject the null hypothesis, you know the probability that you are making the wrong decision.	F
If the experiment were repeated thousands of times, you would obtain a significant result ~ 99% of the time.	F

This assumes the null hypothesis is false, or equivalently, the alternative hypothesis is true. ($p = 0.01$ could be Type I error)



P-Values: Recommendations

- Remember
 - Statistics don't "prove" anything
 - The p-value is not the probability of a hypothesis
 - Unfortunately, we can reject the hypothesis that most p-values are interpreted correctly.

Review Lessons 1-4

Parting Thoughts



Review: Four Lessons

4. A picture is worth 1,000 words.
Examine your data before embarking on analysis.

3. Not all observations are independent.
Recognize dependencies in your data
Use methods that account for the dependencies

2. Are 50% of us really above average?
Difference between mean and median
Don't look at either in isolation (see #1)
SD describes population; SE describes estimate from sample

1. What the **** is a p-value?
Not a probability of the null/alternative hypothesis



Statistically Speaking ...

What's next?

- Tuesday, October 11 **Statistical Considerations for Sex Inclusion in Basic Science Research**
Denise M. Scholtens, PhD, Associate Professor, Division of Biostatistics Associate Director, Department of Preventive Medicine
- Friday, October 14 **The Impact of Other Factors: Confounding, Mediation, and Effect Modification** Amy Yang, MS, Sr. Statistical Analyst, Division of Biostatistics, Department of Preventive Medicine
- Tuesday, October 18 **Statistical Power and Sample Size: What You Need and How Much**
Mary Kwasny, ScD, Associate Professor, Division of Biostatistics, Department of Preventive Medicine
- Friday, October 21 **Clinical Trials: Highlights from Design to Conduct** Masha Kocherginsky, PhD, Associate Professor, Division of Biostatistics, Department of Preventive Medicine
- Tuesday, October 25 **Finding Signals in Big Data** Kwang-Youn A. Kim, PhD, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine
- Friday, October 28 **Enhancing Rigor and Transparency in Research: Adopting Tools that Support Reproducible Research** Leah J. Welty, PhD, BCC Director, Associate Professor, Division of Biostatistics, Department of Preventive Medicine

All lectures will be held from noon to 1 pm in Hughes 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