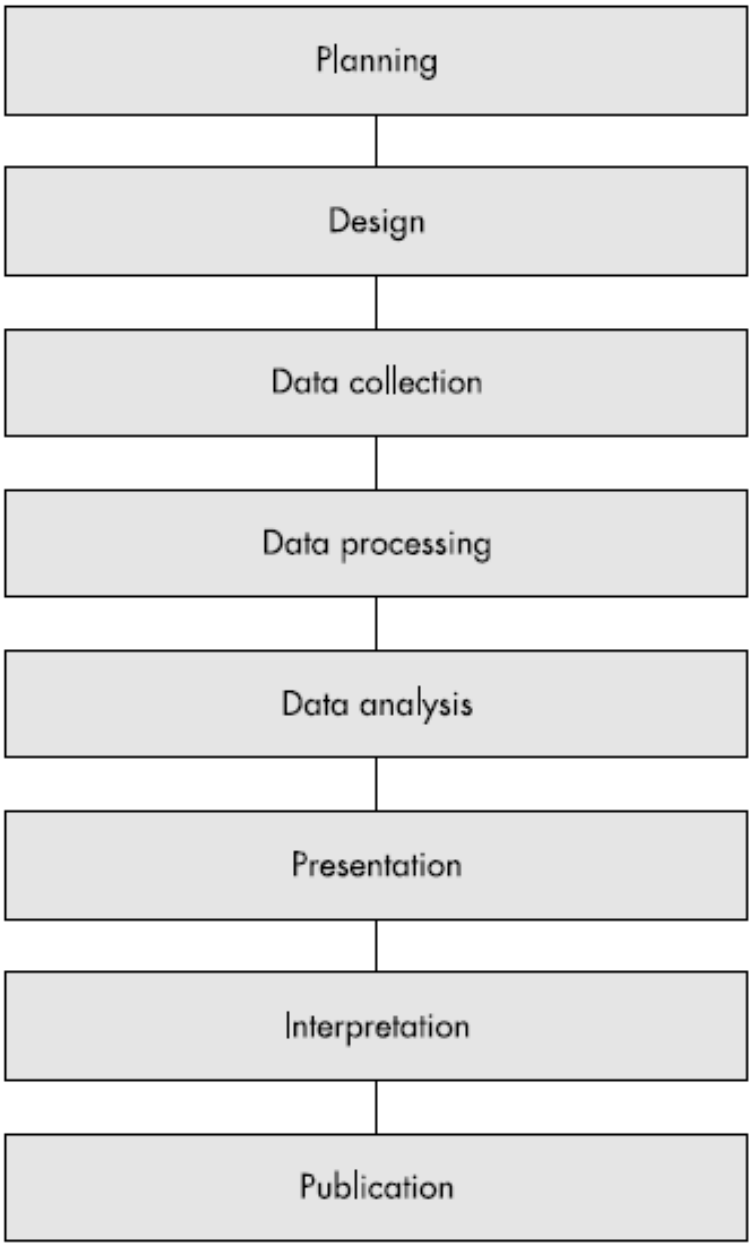
# Basic Statistical Considerations for Quantitative Educational Studies

**Chunfa “Charles”Jie, PhD**

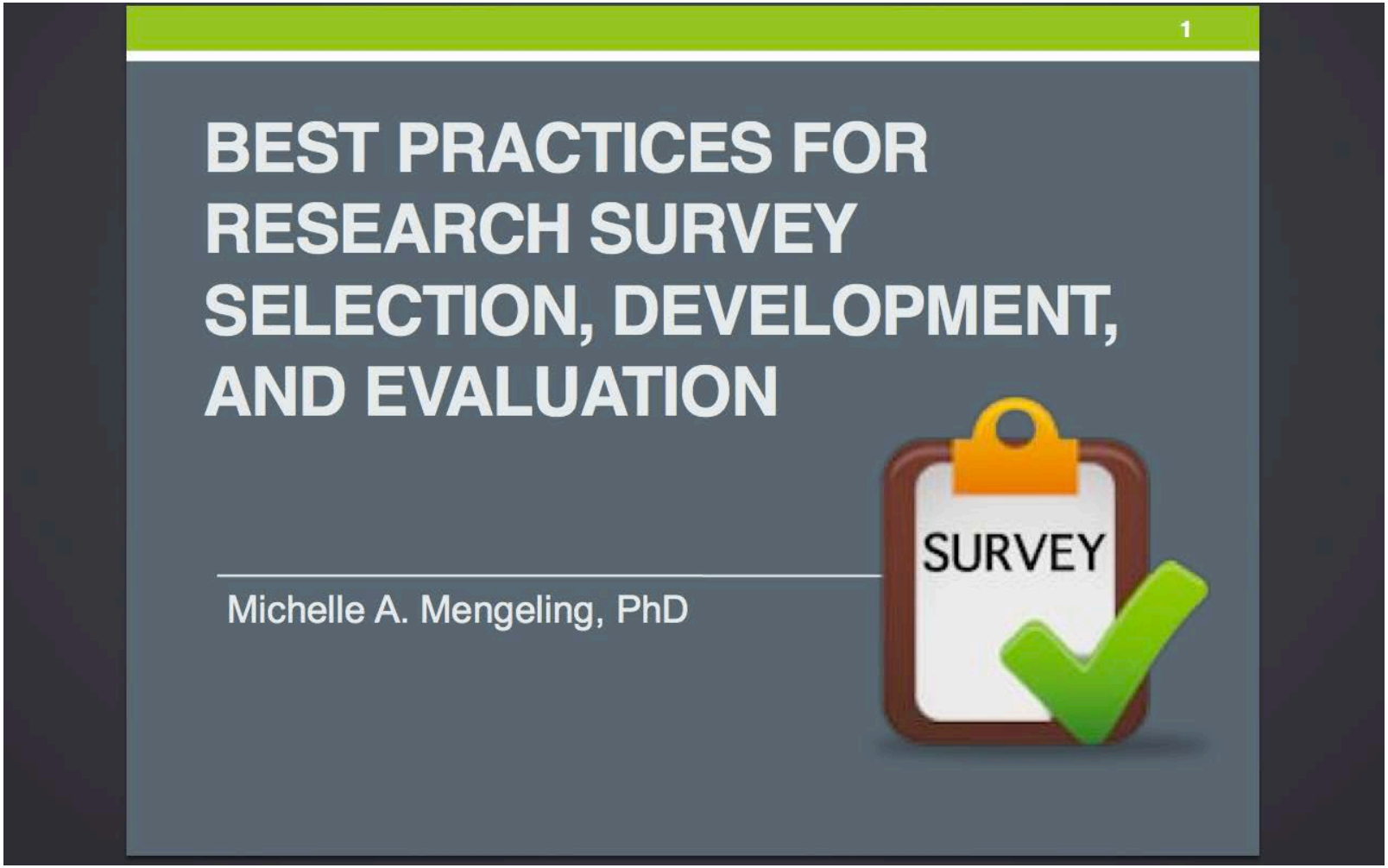
Dr. Jie indicated he has no financial relationships to disclose relevant to the content of this CME activity.

# Statistics




S Carley, F Lecky

Health survey research ---
A starting point to show the roles and considerations of statistics



**A simple clinical case:** Mercy Physicians

**Goal** : Closing the loop ---to improve data sharing between specialists and the PCPs

- "Provided the patient with a pamphlet (PHC passport) that has our office details, along with written instructions to please forward all medical records pertinent to patient care."
- "our hope is to close the loop more efficiently."

- measurement tool:

a physician survey before and 6 months after the patients have started using the pamphlet to ascertain whether this has resulted in decreased "wasted visits."

# Data of two questions from the survey

Beneficial to continue PHC passport?

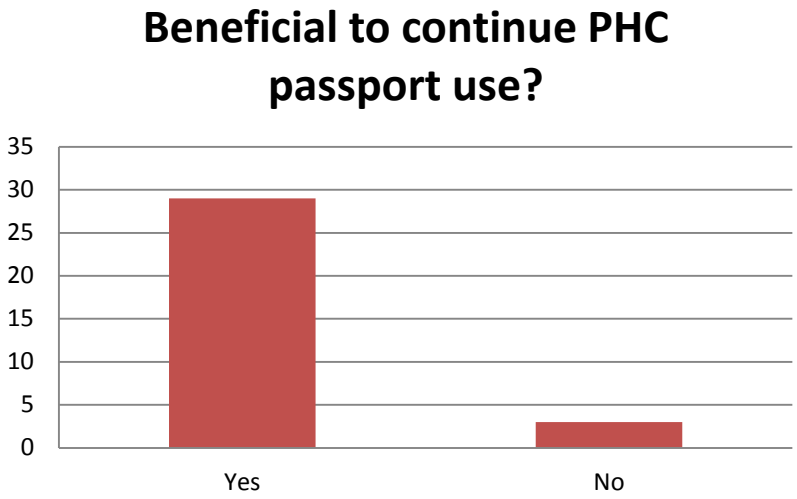
| | |
|---|---|
| Yes | 29 |
| No | 3 |

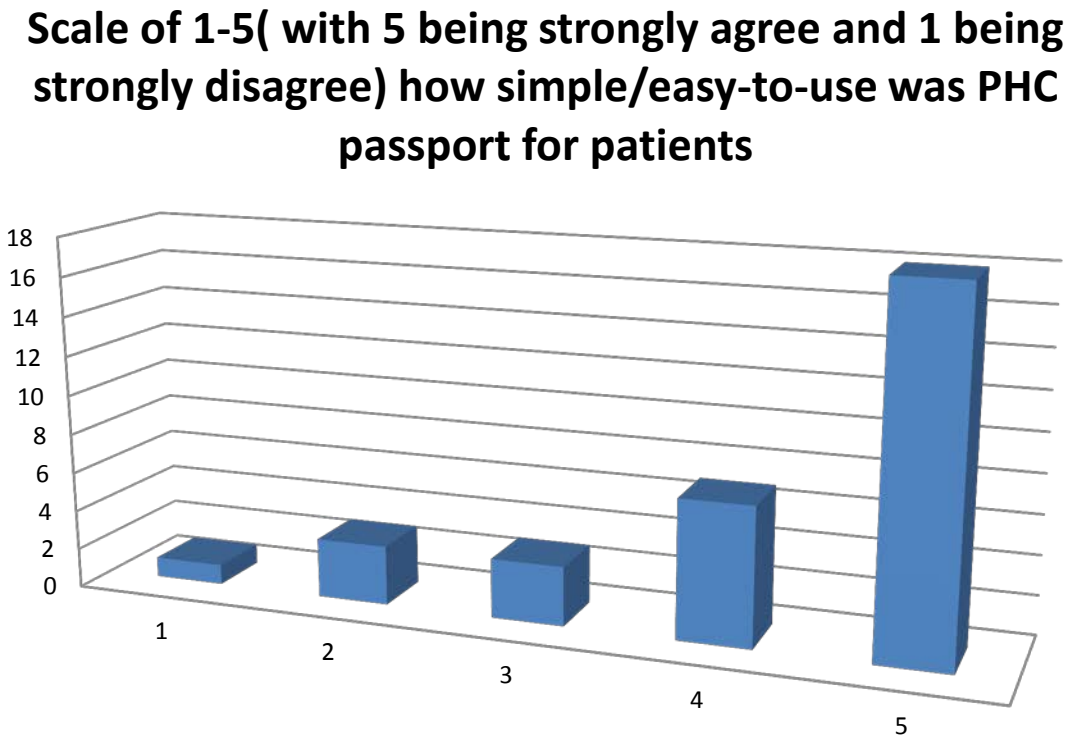
Scale of 1-5( with 5 being strongly agree and 1 being strongly disagree)

how simple/easy-to-use was PHC passport for patients

| | |
|---|---|
| 1 | 1 |
| 2 | 3 |
| 3 | 3 |
| 4 | 7 |
| 5 | 18 |

# Descriptive statistics:



The inferential statistics we can do at most with such data:

**analysis only for each variable**

1. Beneficial to continue PHC passport?

The proportion of the "yes" response is estimated to be 0.906 with a 95% confidence interval of [0.75, 0.98].

2. how simple/easy-to-use was PHC passport for patients?

Scale of 1-5( with 5 being strongly agree and 1 being strongly disagree)

There is sufficient evidence to show that the proportions of the five categories (1-5) are different. (p-value=0.0005)

**Question**: Is there any connection between the two survey Questions? Can the data answer this question?

**How can we make better use of research data?**

Case study: research by Dr. Rebecca Shaw

A survey of women's health care conducted in a rural area in the Dominican Republic by DMU medical students

- 53 women approached randomly at the local clinics
- Survey consisted of 11 questions

| *Questions:* |
| --- |
| **Age** |
| **Gravida** |
| **Birthing location** |
| **Most common birthing location** |
| **Contraceptives (Y/N)** |
| **Type** |
| **Interest in Family Plan (Y/N)** |
| **Short or long FP** |
| **Where receive family planning** |
| **Zika** |
| **Prenatal check-ups** |

| | | | | |
|---|---|---|---|---|
| **Age** | 42 | 22 | 30 | 25 |
| **Gravida** | 5 | 1 | 3 | 2 |
| **Birthing location** | Monte Cristi | Monte Cristi | Santiago Rodriquze | Hospital Mao |
| **Most common birthing location** | Monte Cristi Hospital | Clinic | Monte Cristi Hospital | Monte Cristi Hospital |
| **Contraceptives (Y/N)** | Y | Y | N | Y |
| **Type** | Pills | Pills | | Condoms/Pills/ Injection |
| **Interest in Fam Plan (Y/N)** | N | Y | Y | Y |
| **Short or long FP** | N/A | Long | Long | Short |
| **Where receive family planning** | Hospital | Pharmacy (w/o Rx) | | La Clinica Madre |
| **Zika** | No change | No change | Doesn't know | No change |
| **Prenatal check-ups** | every 15 days from Dr. Garcia and monthly in the hospital | every month in Monte Cristi | every month in Monte Cristi | every 15 days from Dr. Garcia |

**Organize data:**

1. Standardize free text data entries:

e.g. **Birthing Location**
Santiago Rodriquze, Hospital Mao, Moka, Monte Cristi, Hospital Manuel de Luna, San Juan de la Manguna, Monte Crisit Hospital, Santo Domingo, Santiago …

**e.g. Where to receive family planning?**
Hospital, Pharmacy (w/o Rx), La Clinica Madre, Pharmacy, Batey Clinica, …

2. Re-code the variables in a clinically meaningful way:

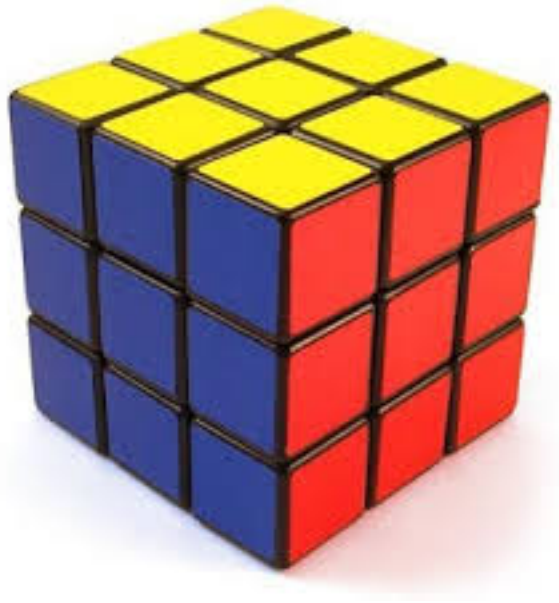
e.g. Size of medical facility may be more important.

**Birthing Location**:

large hospital vs small clinic

3. Collapse the categories of a variable:

cross-tabulation of two variable with many categories will lead to no data points for some combinations of the two variables.



e.g. Gravida

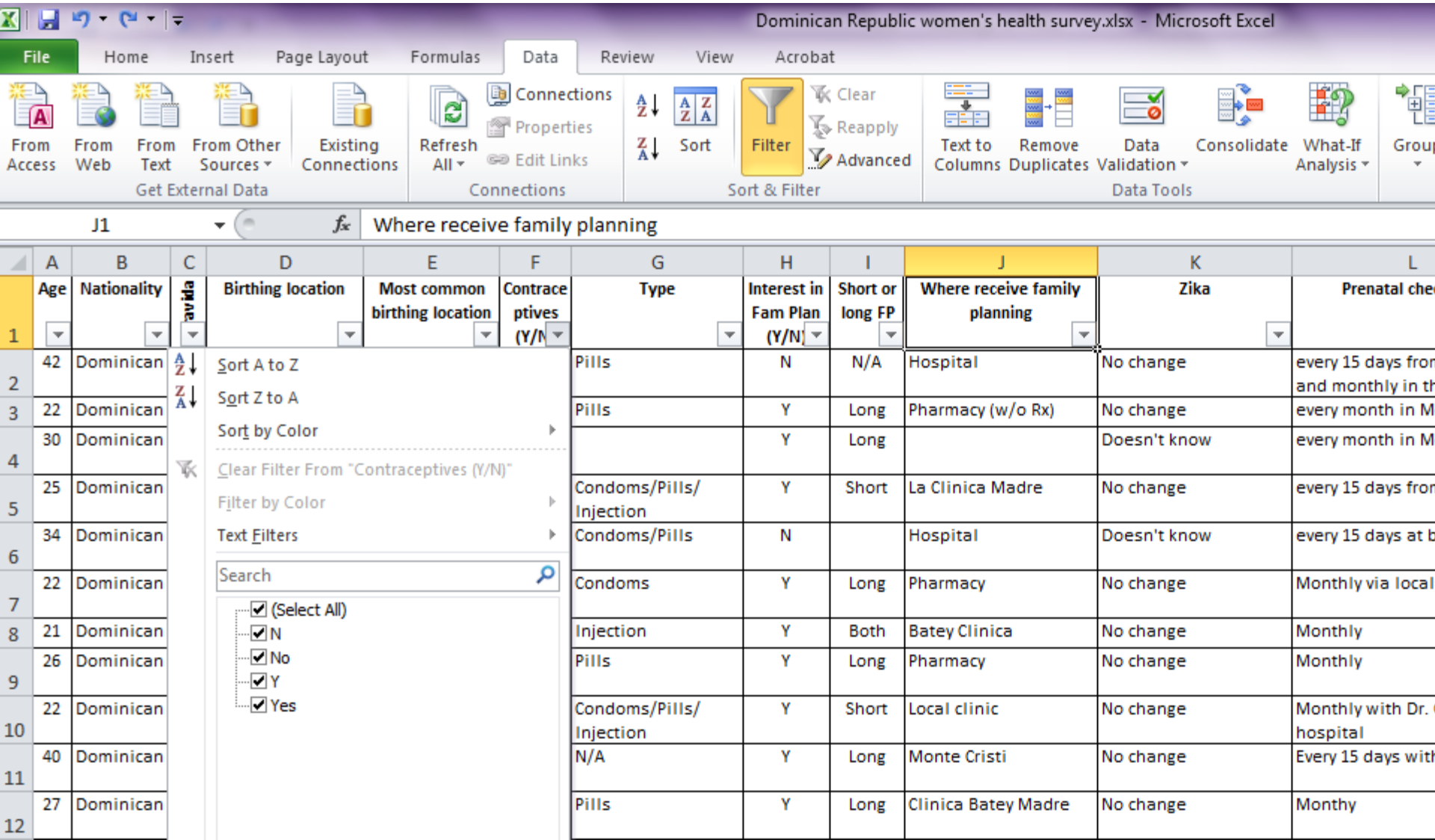
Original values: 0, 1, 2, 3, 4, 5, 6, 11

**Revised coding for Family Size/Gravida**

| | |
|---|---|
| None: | gravida 0 |
| average: | gravida 1-2 |
| Large: | gravida 3+ |

4. Use EXCEL filter/unique function to check the categories and unexpected data entries/errors:

e.g. N/A, blank cell, unsure, ---




| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | Nationality | avida | Birthing location | Most common birthing location | Contraceptives (Y/N | Type | Interest in Fam Plan (Y/N | Short or long FP | Where receive family planning | Zika | Prenatal che |
| 2 | 42 | Dominican | | | | | Pills | N | N/A | Hospital | No change | every 15 days fro and monthly in th |
| 3 | 22 | Dominican | | | | | Pills | Y | Long | Pharmacy (w/o Rx) | No change | every month in M |
| 4 | 30 | Dominican | | | | | | Y | Long | | Doesn't know | every month in M |
| 5 | 25 | Dominican | | | | | Condoms/Pills/ Injection | Y | Short | La Clinica Madre | No change | every 15 days fro |
| 6 | 34 | Dominican | | | | | Condoms/Pills | N | | Hospital | Doesn't know | every 15 days at b |
| 7 | 22 | Dominican | | | | | Condoms | Y | Long | Pharmacy | No change | Monthly via local |
| 8 | 21 | Dominican | | | | | Injection | Y | Both | Batey Clinica | No change | Monthly |
| 9 | 26 | Dominican | | | | | Pills | Y | Long | Pharmacy | No change | Monthly |
| 10 | 22 | Dominican | | | | | Condoms/Pills/ Injection | Y | Short | Local clinic | No change | Monthly with Dr. hospital |
| 11 | 40 | Dominican | | | | | N/A | Y | Long | Monte Cristi | No change | Every 15 days with |
| 12 | 27 | Dominican | | | | | Pills | Y | Long | Clinica Batey Madre | No change | Monthy |

**Start with 2 or 3 categories for a variable in a initial pilot study**.

data file column headers with one or two words saved in the tab delimited text (txt) or coma separated (csv) format.

| Batey | Fertility | Family. Size | Contraceptives | Usage | Interest. FP | FP.S.L | Zika.Ed | Zika. Aware | Change.FP |
|---|---|---|---|---|---|---|---|---|---|
| Isabel | late | large | Y | 1 | N | N/A | 0 | 1 | 0 |
| Isabel | early | average | Y | 1 | Y | Long | 0 | 0 | |
| Isabel | early | large | N | 0 | Y | Long | 0 | 0 | |
| Isabel | early | average | Y | 1 | Y | Short | 0 | 1 | 0 |
| Isabel | late | large | Y | 1 | N | N/A | 0 | 0 | |
| Isabel | early | large | Y | 1 | Y | Long | 0 | 1 | 0 |
| Isabel | early | none | Y | 1 | Y | | 0 | 1 | 0 |
| Isabel | early | average | Y | 1 | Y | Long | 0 | 1 | 0 |
| Isabel | early | none | Y | 1 | Y | Short | 0 | 1 | 0 |

## Data analysis:

Analysis can be performed beyond the level of each individual Variable.

1) Association analysis between the **nominal** variables:

Nominal variable: categorical variable without ranked levels

e.g. Contraceptive: Yes or no

desire to change family planning choices due to information about zika virus (Change.fp): Yes or No

Zika Education (Zika.Ed): Yes or No

**Zika.Ed_vs_Change.FP**

| | No | Yes |
|---|---|---|
| No | 19 | 1 |
| Yes | 11 | 17 |

p-value= 0.0004  Contigency Coeffiicent=0.493

**Contraceptives_vs_Change.FP**

| | No | Yes |
|---|---|---|
| No | 12 | 1 |
| Yes | 18 | 17 |

p-value= 0.0154  Contigency Coeffiicent= 0.351

2) Association analysis between the **ordinal** variables:

Ordinal variable: categorical variable with ranked levels

e.g. Fertility: early (18-30), late (31-50), post (51+)

Family Size:

None: gravida 0

average: gravida 1-2

Large: gravida 3+

Fertility_vs_Family.Size

| | none | average | large |
|---|---|---|---|
| early | 9 | 16 | 6 |
| late | 0 | 3 | 11 |
| post | 0 | 1 | 7 |

p-value= 0.000008

Kendall's rank correlation tau =0.57

3) Association analysis between the ***ordinal*** and ***nominal*** variables:

***Interest in Family Planning***: Yes (Y) and No (N)
***Fertility***: early (18-30), late (31-50), post (51+)

| | | ***Fertility*** | | |
|---|---|---|---|---|
| | | early | late | post |
| ***Interest in Family Planning***: | N | 3 | 7 | 7 |
| | Y | 27 | 7 | 1 |

| | | ***Fertility*** | | |
|---|---|---|---|---|
| | | early | late | post |
| ***Interest in Family Planning***: | N | 0.100 | 0.500 | 0.875 |
| | Y | 0.900 | 0.500 | 0.125 |

***P-value=0.00001 (Armitage Trend test)***

***Contraceptives***: Yes (Y) and No (N)

***Fertility***

| ***Contraceptives***: | early | late | post |
|---|---|---|---|
| N | 4 | 7 | 5 |
| Y | 27 | 7 | 3 |

***Fertility***

| ***Contraceptives***: | early | late | post |
|---|---|---|---|
| N | 0.129 | 0.500 | 0.625 |
| Y | 0.871 | 0.500 | 0.375 |

***P-value=0.001 (Armitage Trend test)***

| paired_variables | p.value | contingency_ coefficient |
|---|---|---|
| Batey_vs_Zika.Ed | 0.0002 | 0.707 |
| Batey_vs_Zika.Aware | 0.0246 | 0.393 |
| Batey_vs_Change.FP | 0.001 | 0.5 |
| Fertility_vs_Family.Size | 0.0006 | 0.537 |
| Fertility_vs_Contraceptives | 0.002 | 0.414 |
| Fertility_vs_Usage | 0.0328 | 0.424 |
| Fertility_vs_Interest.FP | 0.0002 | 0.526 |
| Family.Size_vs_Interest.FP | 0.0092 | 0.391 |
| Contraceptives_vs_Usage | 0.0002 | 0.707 |
| Contraceptives_vs_Interest.FP | 0.0108 | 0.348 |
| Contraceptives_vs_Change.FP | 0.0162 | 0.351 |
| Usage_vs_Interest.FP | 0.0124 | 0.364 |
| Usage_vs_Change.FP | 0.0034 | 0.4 |
| Zika.Ed_vs_Change.FP | 0.0004 | 0.493 |

4) Generalized Linear mixed-effects regression Modeling analysis to assess the association of one response variable with other covariates.

**Clinical trial typically have one primary endpoint.**

**NIAID: Investigator-Initiated Clinical Trial R01 Implementation Grants**

***One of key requirements: Clear primary and secondary endpoints***

Sample size may need to be increased if there are more than one primary endpoint.

## Zika.Ed_vs_Change.FP

Zika Education (*Zika.Ed*): Yes or No

Desire to change family planning choices due to information about zika virus (*Change.fp*): Yes or No

| | | **Change.FP** | |
|---|---|---|---|
| | | No | Yes |
| **Zika.Ed** | No | 19 | 1 |
| | Yes | 11 | 17 |

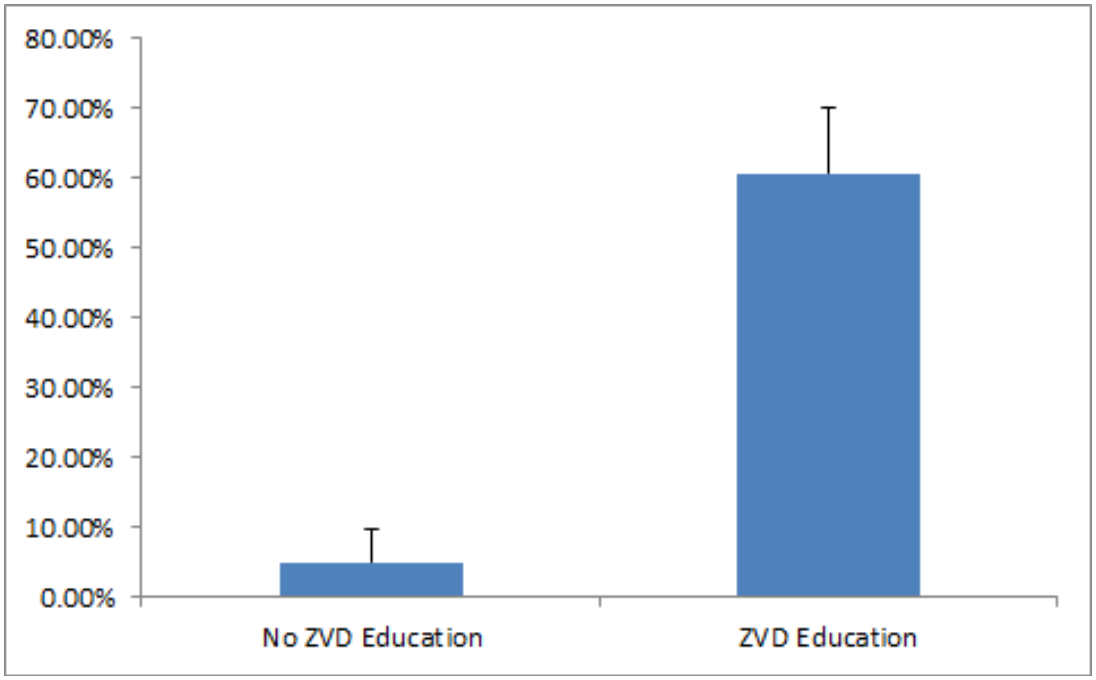
p-value= 0.0004

Contigency Coeffiicent=0.493

| | **Change.FP** | |
|---|---|---|
| | No | Yes |
| No | 0.9500 | *0.0500* |
| Yes | 0.3929 | *0.6071* |

**Comparing the proportions:**

**Influence of Zika Virus Disease (ZVD) Education on Desire to Change Contraceptive**




**Error bar: sqrt(p*(1-Θ)/n) where Θ is the proportion**

**Given the high publicity of Zika virus, we may set the primary endpoint of a future study:**

**Influence of Zika Virus Disease (ZVD) Education on Desire to Change Contraceptive**

**Generation of Research Hypothesis:**

**The Zika Virus Disease (ZVD) Education affects Desire to Change Contraceptive.**

# Planning a study

**Upfront sample size/power calculation and downstream data analysis are affected by**

- Context of the data
- Source of the data
- Types of the data
- Sampling method
- Study design

# Research Hypothesis

- Research hypothesis is the definite statement that there is a relationship (or difference) between variables
- Examples:
  - There is a positive relationship between income and years of education.
  - Teenage pregnancy rates differ among geographical regions.
  - There is a difference between genders and numb times per week that ice cream is consumed.

# Direction of Research Hypotheses

- **Nondirectional research hypotheses**
  - States that there is a difference (or relationship) between groups but doesn't specify the direction

  Example: There is ***a difference*** between genders (men & women) regarding number of times per week that ice cream is consumed.

- **Directional research hypotheses**
  - States that there is a difference (or relationship) between groups AND specifies direction

  Example: The average number of times that men eat ice cream per week is ***greater than*** the average number of times that women each ice cream per week.

# Statistical Way: Hypotheses

- Starting Point:

  Null hypothesis ($H_0$ )= essentially states that there will be no difference or no relationship between your variables

- Examples:
  - There will be **no** relationship between income and years of education.
  - There will be **no** difference in teen pregnancy rates among geographical regions.
  - There is **no** difference between gender and number of times per week that ice cream is consumed.

# Null Hypothesis

Why is it necessary?

- Starting point against which actual outcomes can be measured

*There will be no relationship between income and years of education.*

- You start out your research by assuming that two variables (income & years education) are not related
- Additionally, until demonstrated otherwise, you assume that the relationship is due to chance

- We run statistical analyses to minimize likelihood of relationship or differences being due to chance

Research Hypothesis becomes the alternative hypothesis ($H_a$ or $H_1$ ) in a statistical test setting.

# Example from a BSMS student

- 1. Is there a significant difference between HSA and C1q treated macrophages in engulfment for the WT macrophages and the LAIR-/- macrophages?
  - $H_0$: There is no significant difference in engulfment between HSA and C1q treated macrophages ($u_1 = u_2$)
  - $H_A$: There is a significant difference in engulfment between HAS and/C1q treated macrophages ($u_1 \neq u_2$)
- 2. Is there an overall decrease in engulfment for the LAIR -\- macrophages?
  - $H_0$: There is no decrease in engulfment of the LAIR-\- macrophages compared to WT ($u_1 = u_2$)
  - $H_A$: There is a decrease in engulfment of the LAIR-\- macrophages compared to WT (($u_1 > u_2$)

**Hypothesis-based Sample Size calculations depends on:**

- Data type
- statistical tests ( t-test, ANOVA) relevant to the study design
- Significance level
- Power

**Common scenarios:**

To compare the means of two populations, we also need to know:

- Standard deviation or variance
- Difference

To compare two proportions, we also need to know:

- Proportions estimated from samples

**Example:**
Patients who are discharged following an admission for acute decompensated heart failure (ADHF) are particularly vulnerable and at high risk for readmission in the next 30 days.

The readmission rate is estimated to be 25%.
A new monitoring plan is expected to reduce the rate to 5%.

Let **$\Theta$ defines the two proportions.**

$H_0: \Theta_1 = \Theta_2$ vs $H_a: \Theta_1 \neq \Theta_2$ Two-sided

$$n_1 = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \left[r\theta_1(1-\theta_1) + \theta_2(1-\theta_2)\right]}{\left(\theta_1 - \theta_2\right)^2}$$

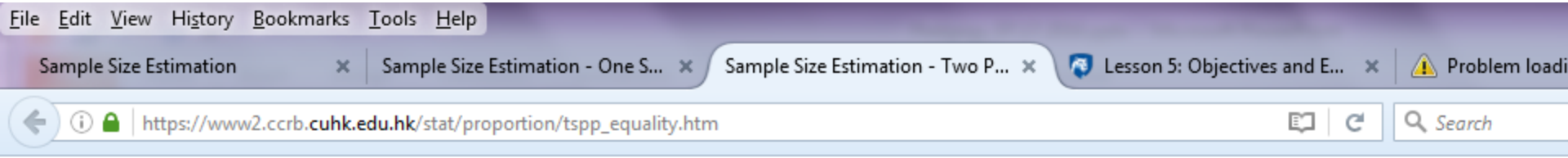
$n_1 = n_2 = 47$

Assuming 25% readmission rate for the conventional group and 5% readmission rate for the group under the new monitoring plan, a total of 94 patients (n=47 per group) would have 80% power to show the difference in readmission rate between the two groups at 5% significance level.

The estimated sample size is typically further adjusted with 15-20% dropout rate.

**Web tool for common types of sample size calculations:**

https://www2.ccrb.cuhk.edu.hk/stat/proportion/tspp_equality.htm




# Sample Size Calculator: Two Parallel-Sample Pro

## Hypothesis: Two-Sided Equality

$H_0: \theta_1 - \theta_2 = 0$ *versus* $H_a: \theta_1 - \theta_2 \neq 0$

**Data Input: (Help) (Example)**

| Input | | | Results | |
|---|---|---|---|---|
| α | 0.05 | Calculate | | |
| β | 0.2 | Reset | $n_1$ | 47 |
| $\theta_1$ | 0.25 | | $n_2$ | 47 |
| $\theta_2$ | 0.05 | | N | 94 |
| r | 1 | | | |

**Note:**

| Variables | Descriptions |
|---|---|
| α | Two-sided significance level |
| 1-β | Power of the test |

**Most hypotheses are like that mentioned above:**

**about "equality":**

Let p defines proportion

$H_{0:}\ \Theta_1 = \Theta_2$ vs $H_{a:}\ \Theta_1 \neq \Theta_2$ Two-sided

Or

$H_{0:}\ \Theta_1 = \Theta_2$ vs $H_{a:}\ \Theta_1 > \Theta_2$ One-sided

Or

$H_{0:}\ \Theta_1 = \Theta_2$ vs $H_{a:}\ \Theta_1 < \Theta_2$ One-sided

For grant proposals, sample size/power calculations based on one-sided hypothesis needs to be justified by prior literature study or pilot data.

**Sample size and power can also be calculated according to Non-inferiority test or Bio-equivalence test:**

Suppose we want to compare an experimental therapy to the standard of care.

**Non-inferiority test:**

The research question in a non-inferiority trial is whether the experimental therapy is not inferior to the standard of care by a predefined margin

**Bio-equivalence test:**

The experimental therapy in an equivalence trial should not be inferior to, nor superior to, the standard of care by a predefined margin.

**Sample Size/power calculations according to confidence interval:**

**Situation: We want to start a new pilot study.**
**No prior or literature information on the standard deviation;**
**No hypothesis of difference between conditions**

Estimate a sample size to ensure there will be enough subjects for a study by specifying a confidence interval.

e.g. 46% medical students have potentially clinical symptoms of depression. (Dr. Lilian Yuan)

N=96 students are needed to ensure 46%±10% (i.e., 36%-56%) of them With 95% confidence to potentially clinical symptoms of depression.

e.g. 46% medical students have potentially clinical symptoms of depression. (Dr. Lilian Yuan)

N=96 students are needed to show that the portion of students with potentially clinical symptoms of depression falls into the 95% confidence interval, **46%±10%** (i.e., 36%-56%) .

N=43 students are needed to show that the portion of students with potentially clinical symptoms of depression falls into the 95% confidence interval, **46%±15%** (i.e., 31%-61%) .

The estimated sample size is typically further adjusted with 15-20% dropout rate.

# Should we pool specimens/samples?

If individual samples are enough,  don't pool them.

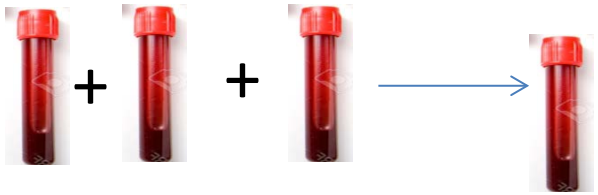
With sample pooing, we will estimate the mean, but  not the standard  deviation or variance.

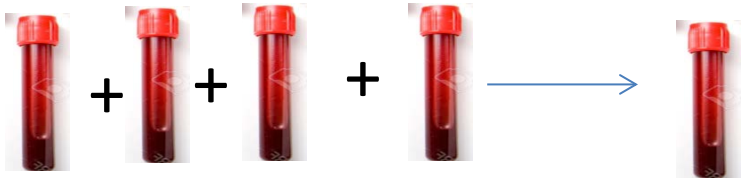
Patient 1 Patient 2

With individual measurements, we will  be able to estimate both the mean and  the standard Deviation or variance.

Pooling is necessary:

e.g. Not enough specimen for each sample

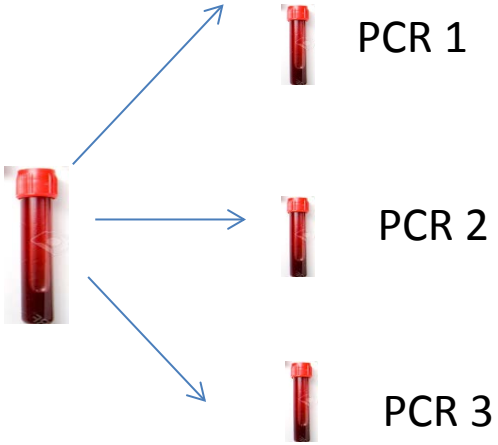




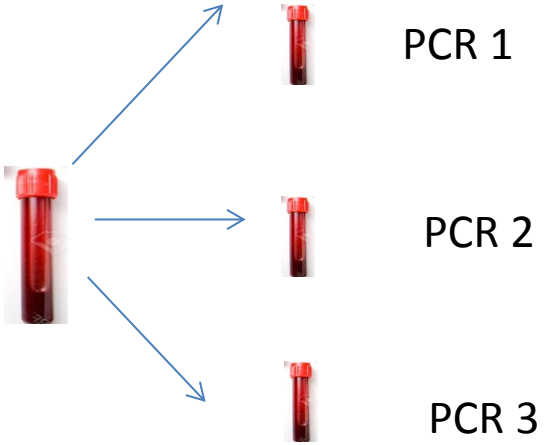
**Better**

**Pooling with an equal number of subjects for each sample has more power.**

# Measure two single samples each multiple times:




Control
Subject 1




patient 1

It is not right to use the three measurement (PCR) values of the same samples to do a t-test. Wrong data variation estimation.

# Log Transformation of Ratio Data

**Range of Ratio data:**

consider rations $y = p/q$ where p and q are both positives in practice

between 0 and infinity

Right skewed

**Log transformation of Ratio data:**

Make data more symmetric and reduce spread

monotone transformation

ratios on straight scale:

$x = 1.8 \quad < \quad y = 18$

ratios on log scale

$\log10(1.8) = 0.255 \quad < \quad \log10(18) = 1.255$

# Background of an example: (Dr. Lauren Ulmer from Mercy Hospital)

BUN/Cr: the **BUN-to-creatinine ratio** is the ratio of two serum laboratory values, the blood urea nitrogen (BUN) (mg/dL) and serum creatinine (Cr) (mg/dL)

The **ratio** of **BUN** to **creatinine** is usually between 10:1 and 20:1. An increased **ratio** may be due to a condition that causes a decrease in the flow of blood to the kidneys, such as congestive heart failure or dehydration

a prospective study :

patients were evaluated to test the diagnostic utility of BUN/Cr ratio as a reliable discriminating factor for upper GI bleeding from lower GI bleeding.

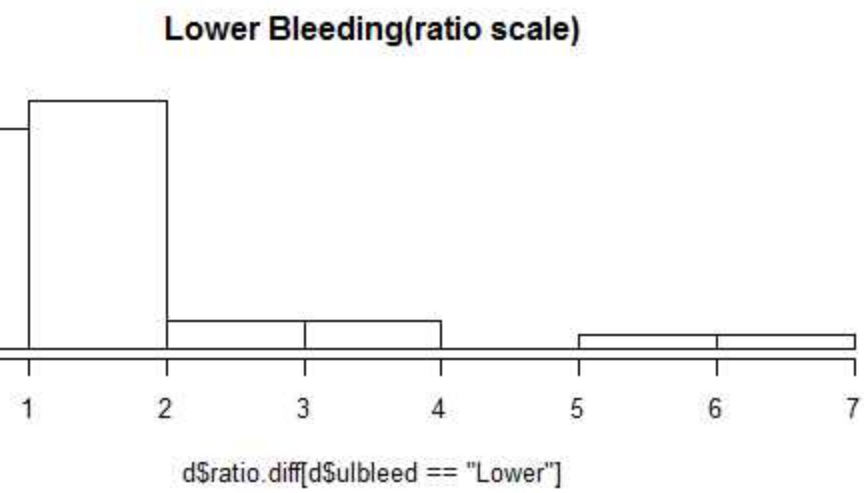
Data Metric of interest: Prior.BUN.Cr.Ratio/admit. PBUN.Cr.Ratio

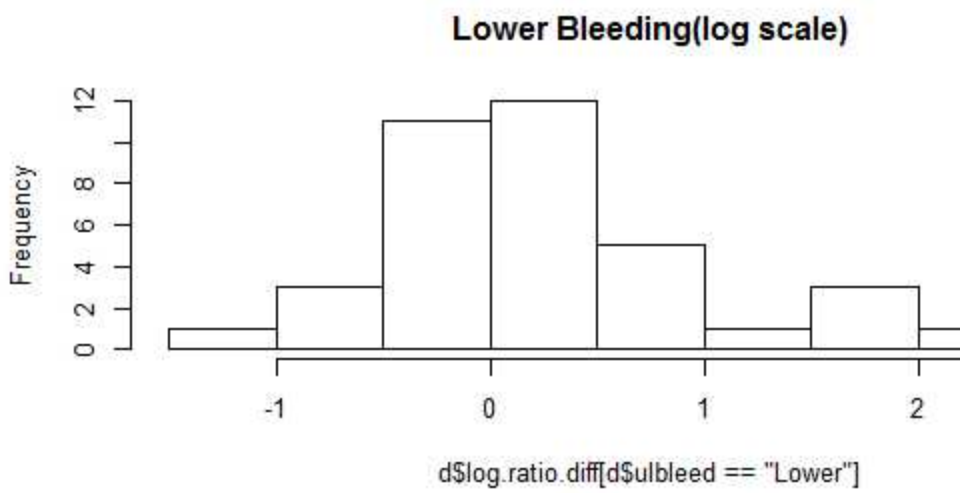
Lower bleeding patients: 38

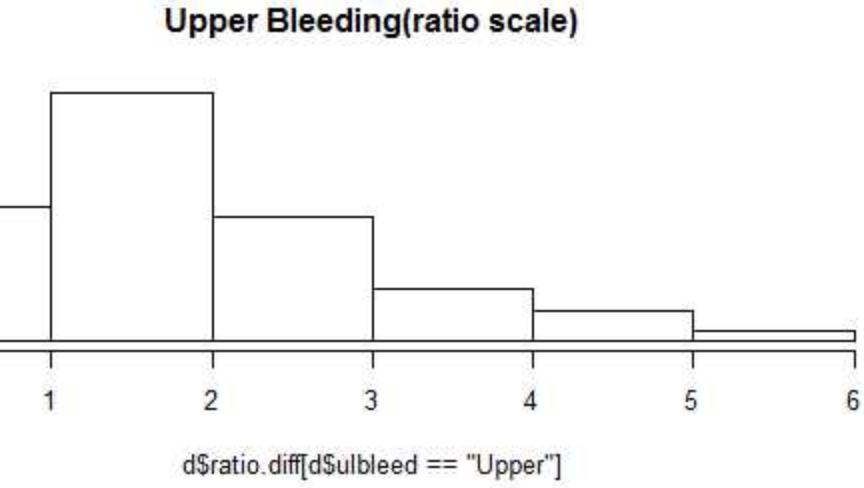
Upper bleeding patients: 58

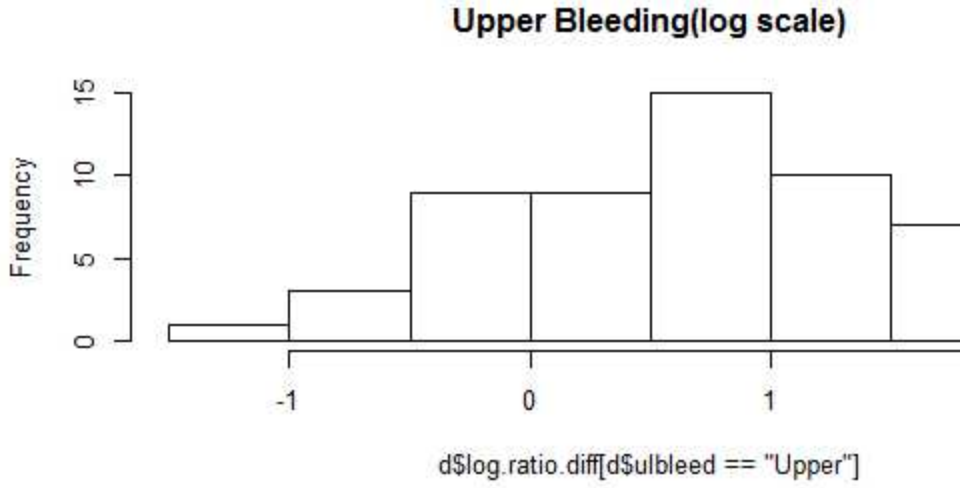
**ratio**

**Log ratio**



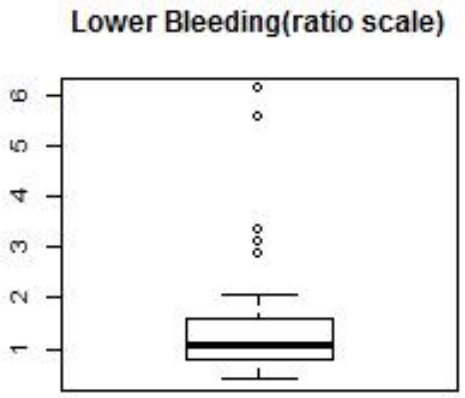



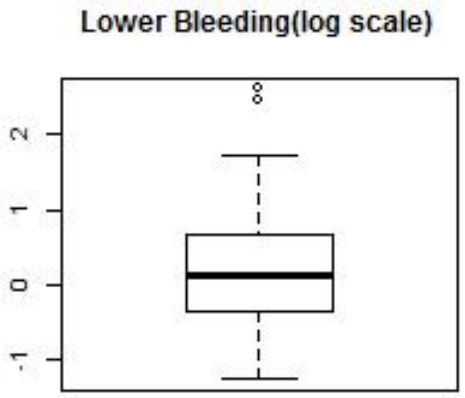



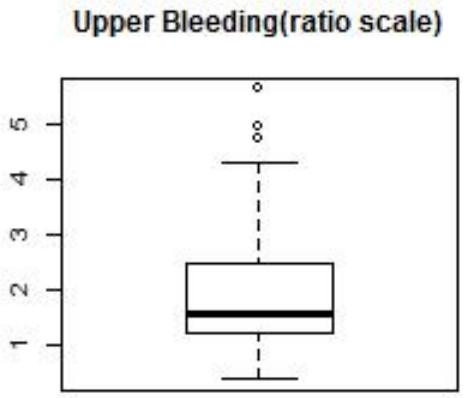



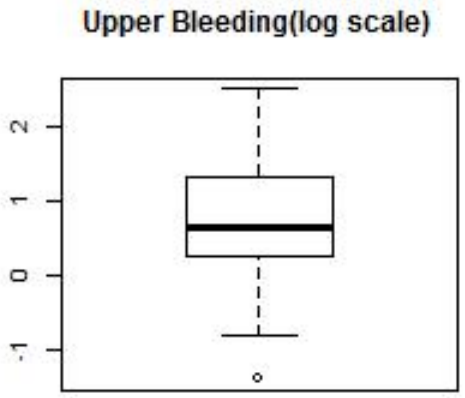
**ratio**

**Log ratio**



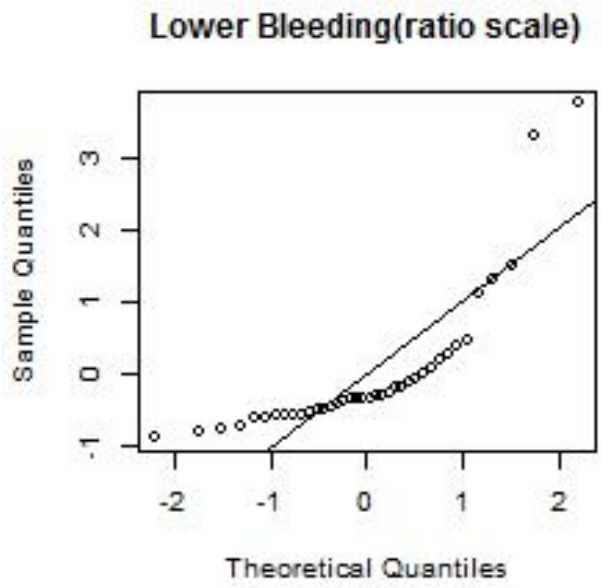



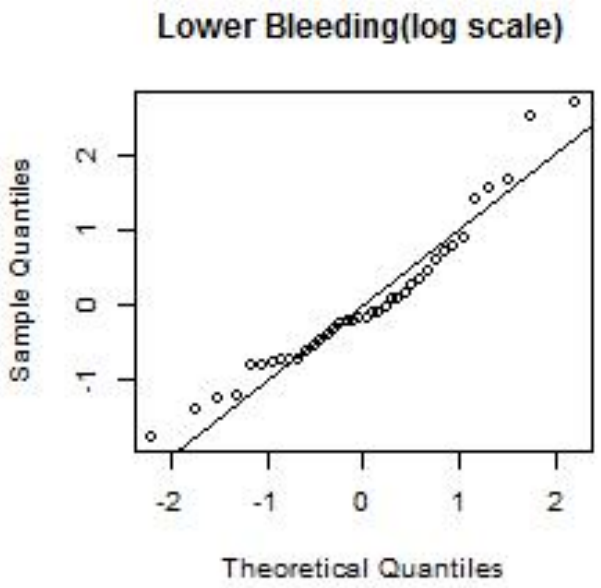



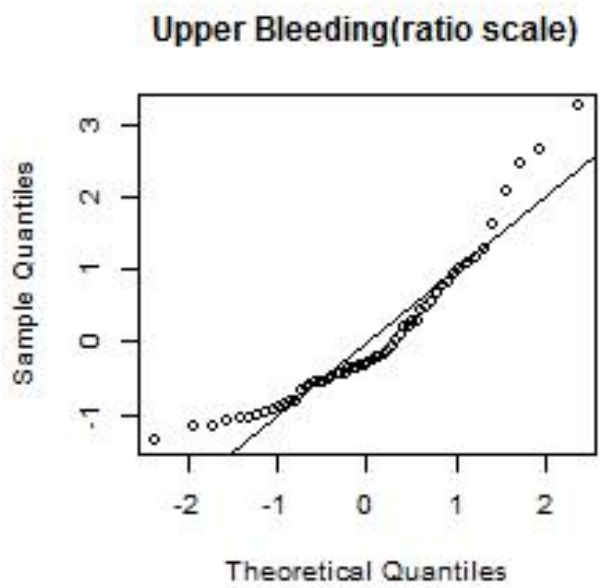



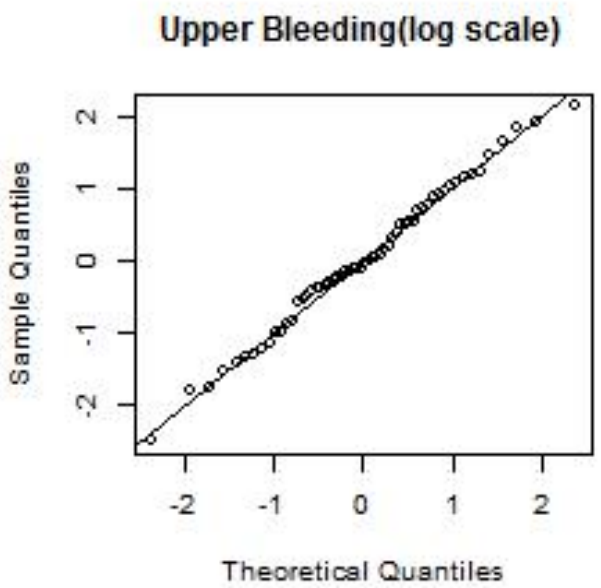
**ratio**

**Log ratio**



# Which analysis result is more reliable:

| | lower | Upper | p-value of the difference |
|---|---|---|---|
| log.ratio. | | | |
| difference | "0.28 +/- 0.86“ | "0.71 +/- 0.84" | "0.0174" |
| ratio. | | | |
| diff | "1.5 +/- 1.24" | "1.93 +/- 1.14" | "0.0867" |

**Practice with the excel data,**
**log-transformation_t-test_demo.xlsx**
**EXCEL function =log10()**

# Angular Transformation of Percent Data

Also known as Arcsine Sqare Root transformation

Applicable to Percent data from binary counting Process

When nearly all values in the data lie between 0.3 and 0.7, there is no need for such transformation.

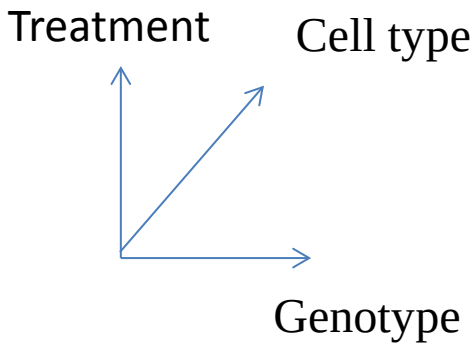
# Background of an example: (Dr. Suzane Bohlson)

Experiment:

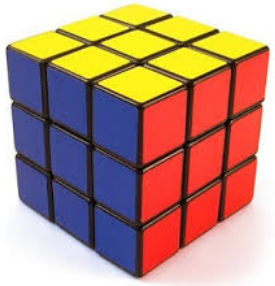
Genotype: Mutant (mt) and Wild Type
Cell Type: Control cell, Apoptotic cells
Treatment: C1q treatment, H treatment

Batch of experiments: 1, 2, 3

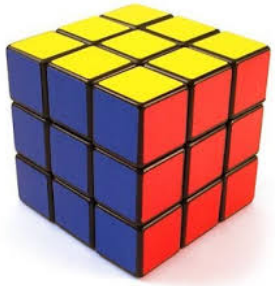
**factorial design 2x2x2x3**



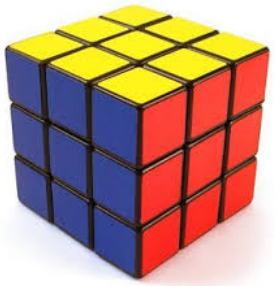

Batch 1



Batch 2

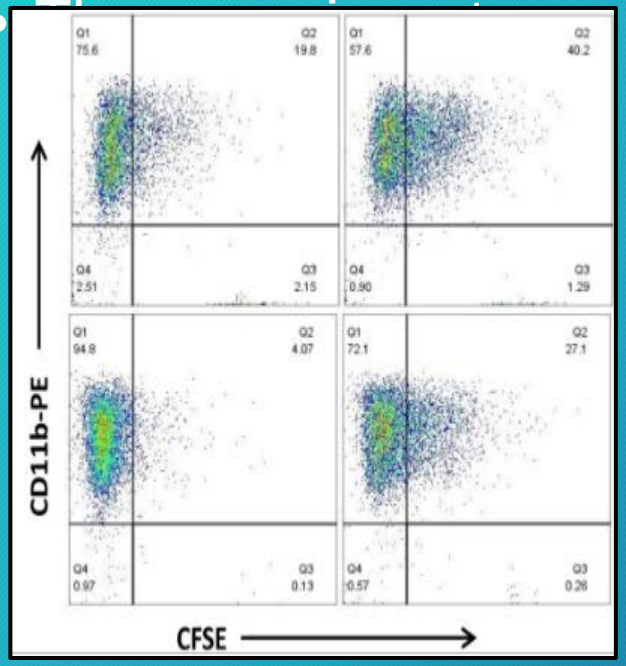


Batch 3



# Data Collection (by a BSMS graduate student)

- Using flow cytometry, percent engulfment could be determined
- repeated three times for

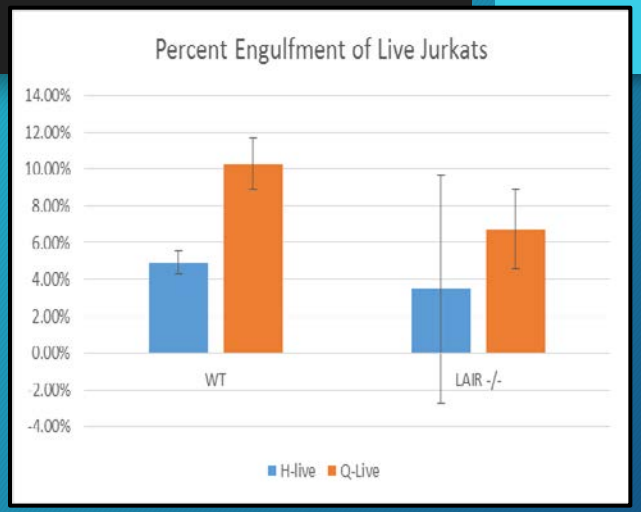


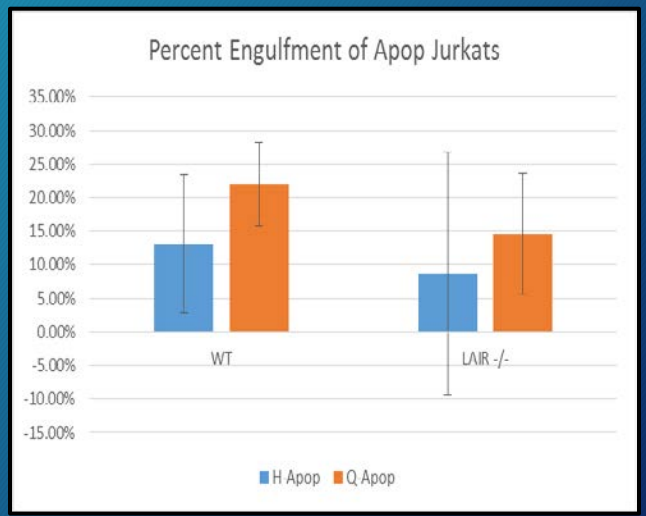

H
Q

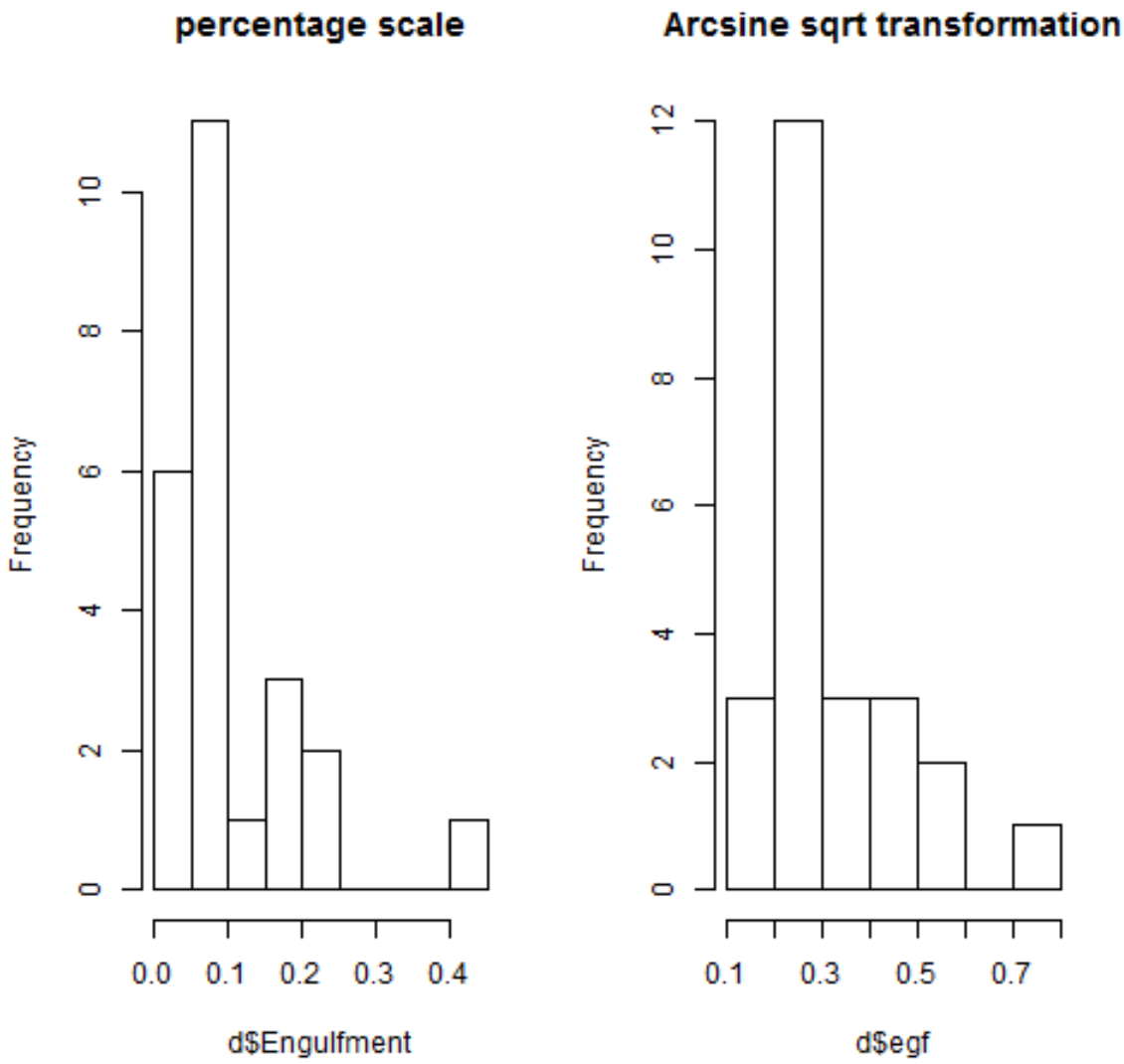
$$\frac{Q_2}{Q_2 + Q_1} = \%\ engulfment$$

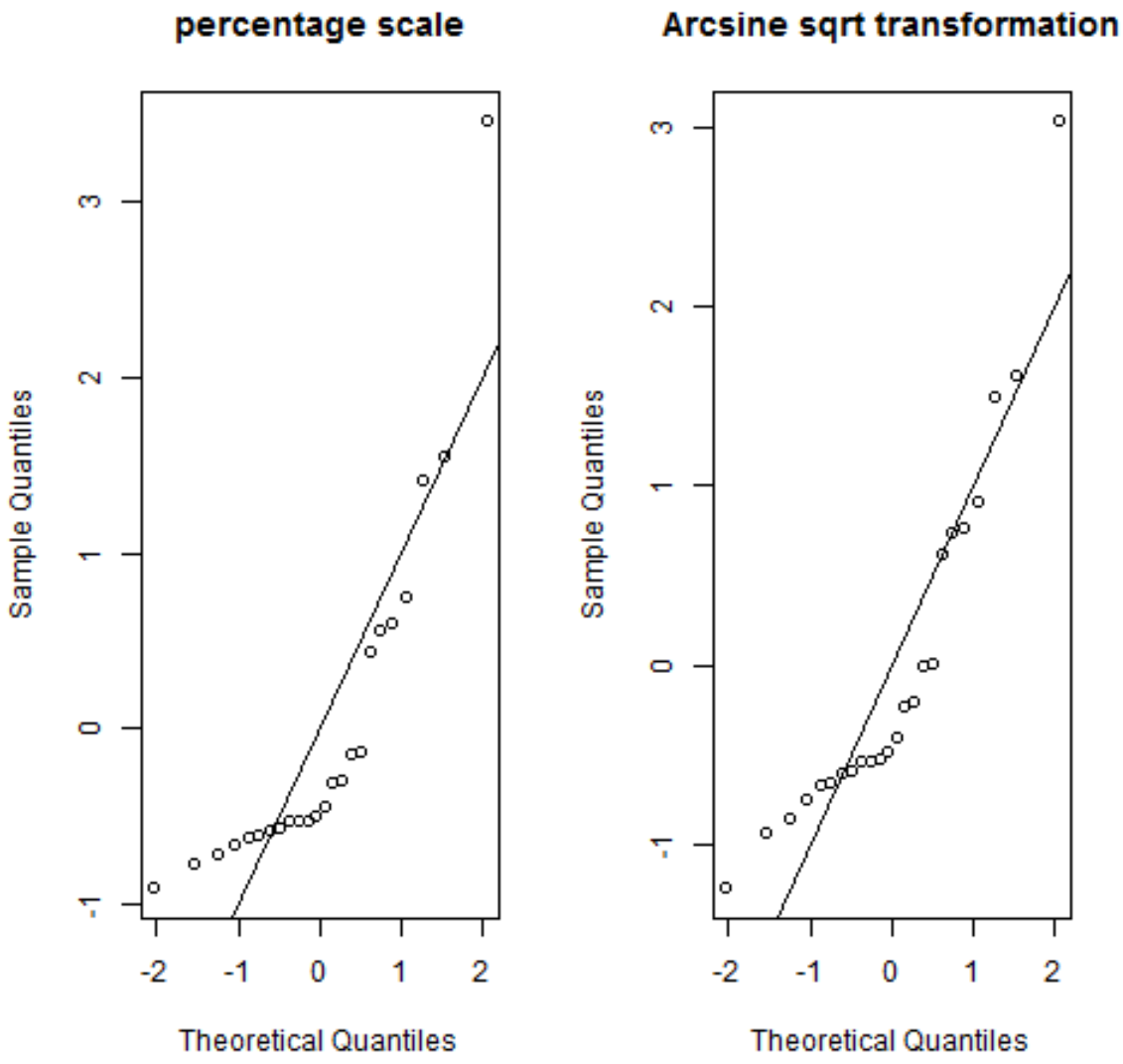
# Data (by a BSMS graduate student)

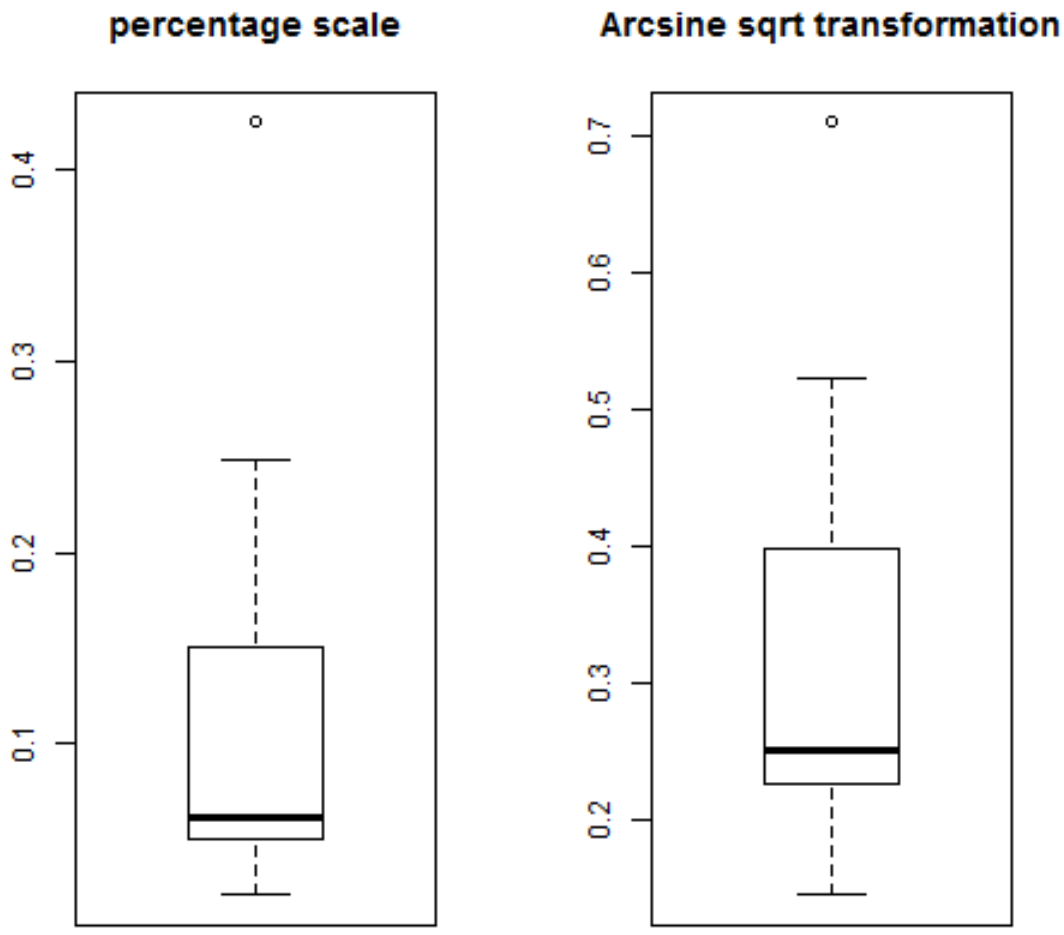
| | H-live | Q-live | H-apop | Q-apop |
|---|---|---|---|---|
| WT | 5.60% | 17.40% | 24.90% | 42.50% |
| LAIR-/- | 4.90% | 9.20% | 15.67% | 23.60% |
| | | | | |
| | | | | |
| | H-Live | Q-Live | H-Apop | Q-Apop |
| WT | 4.40% | 5.90% | 9.30% | 16.00% |
| LAIR -/- | 2.10% | 5.70% | 6.40% | 14.50% |
| | | | | |
| | | | | |
| | H- Live | Q-live | H-apop | Q-apop |
| WT | 4.80% | 7.60% | 5.18% | 7.78% |
| LAIR -/- | 3.40% | 5.30% | 3.85% | 5.59% |




percentage scale
Frequency
0
2
4
6
8
10
0.0
0.1
0.2
0.3
0.4
d$Engulfment
Arcsine sqrt transformation
Frequency
0
2
4
6
8
10
12
0.1
0.3
0.5
0.7
d$egf


percentage scale
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
3
Arcsine sqrt transformation
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2
3


percentage scale
0.1
0.2
0.3
0.4
Arcsine sqrt transformation
0.2
0.3
0.4
0.5
0.6
0.7

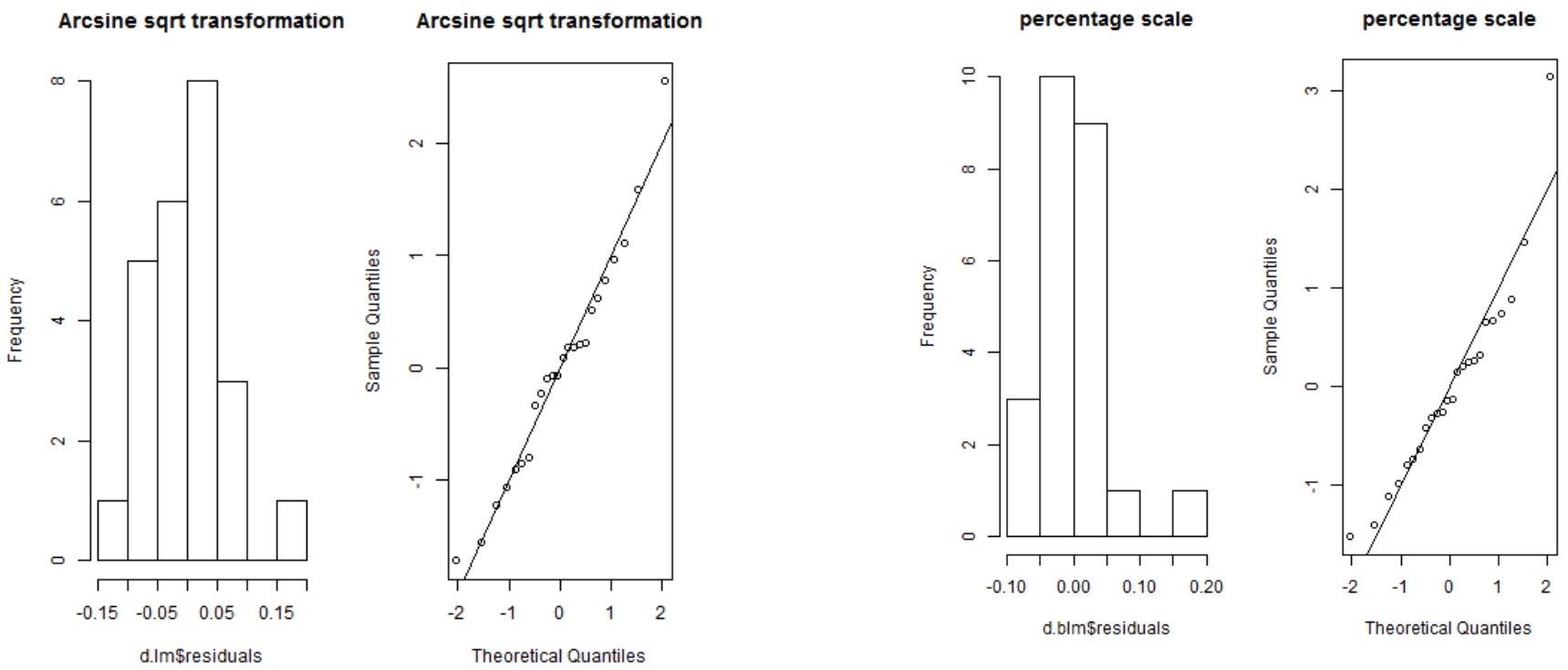
## ANOVA analysis based on a linear model on the transformed data:

| | difference | Std. Error | t value | P-value |
|---|---|---|---|---|
| KO-WT | -0.06332 | 0.02724 | -2.324 | 0.032016 * |
| Q-H | 0.09417 | 0.02724 | 3.457 | 0.002815 ** |
| Live-A | -0.12512 | 0.02724 | -4.593 | 0.000226 *** |
| batch2-batch1 | -0.14205 | 0.03336 | -4.257 | 0.000474 *** |
| batch3 –batch1 | -0.18513 | 0.03336 | -5.548 | 2.88e-05 *** |

## ANOVA analysis based on a linear model on the original percent data:

| | difference | Std. Error | t value | P-value |
|---|---|---|---|---|
| KO-WT | -0.04263 | 0.02254 | -1.891 | 0.074766 . |
| Q-H | 0.05881 | 0.02254 | 2.610 | 0.017738 * |
| Live-Apoptotic | -0.08248 | 0.02254 | -3.660 | 0.001792 ** |
| Batch2-batch1 | -0.09934 | 0.02760 | -3.599 | 0.002051 ** |
| Batch3-batch1 | -0.12534 | 0.02760 | -4.541 | 0.000253 *** |

# Which analysis is more reliable?
# Let's look at the residuals after the two ways of model fitting.



# Comparison of Error Bars between Standard Deviation and Standard Errors

From a reviewer of a manuscript:

“Are the statistical methods used valid?

No

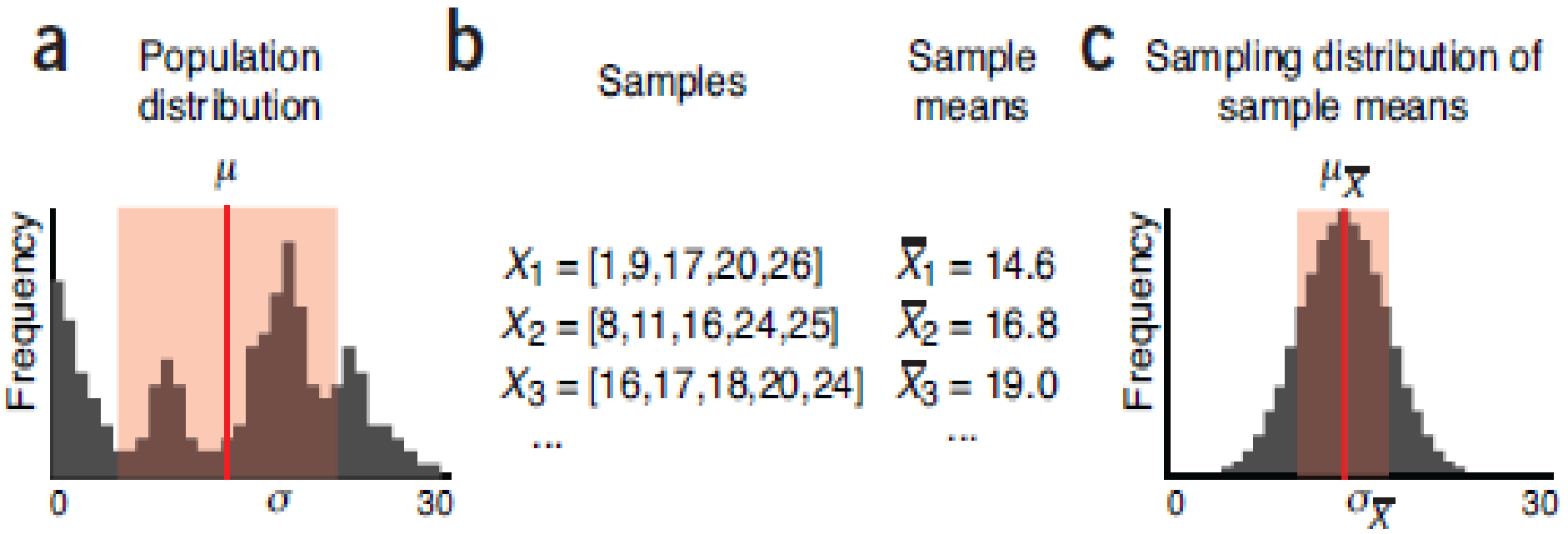
**The standard error mean was used instead of standard deviation**. I would request an advise from a statistician to validate the method for data comparison.”

- The online author guideline of Nature specifies: “Graphs should include clearly labelled error bars. Authors must state whether a number that follows the ± sign is a standard error (s.e.m.) or a standard deviation (s.d.).”

- According to the Nature Method article below, in 2012, 49% of the error bars in the publications of Nature Methods were based on standard error of mean while 45% used standard deviation.

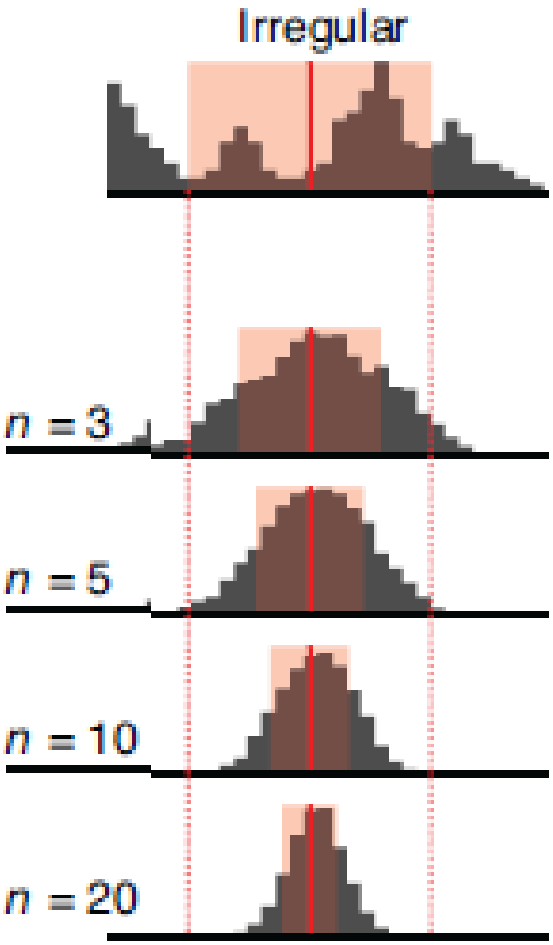
*Altman N. and Krzywinski M. (2013) Points of Significance: Error Bars. Nature Methods, Vol 10 No. 10, 921-922*

**Standard error of the mean depends on the sample size as shown below:**



**Standard error of the mean depends on the sample size as shown below:**

*As n increases, the distribution of sample mean* $\overline{X}$ *has smaller spread.*




Martin Krzywinski & Naomi Altman

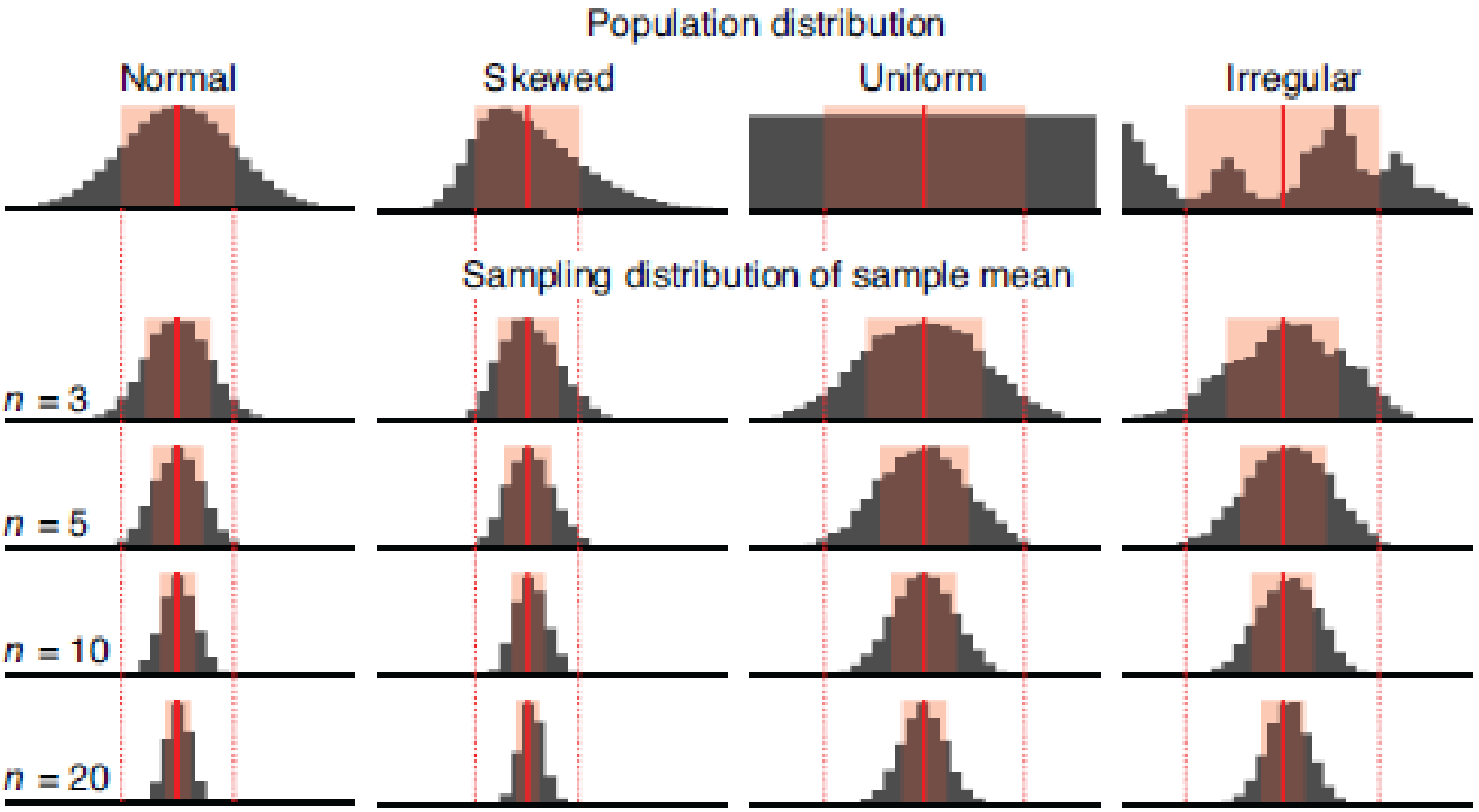
- **The sample standard deviation (*s.d.*) from a particular data set estimates the spread of the population data.**

- **The standard error of mean (*s.e.m*) reflects the uncertainty in the mean estimated from a sample.**

$$s.e.m = \frac{s}{\sqrt{n}}$$

1) *The sample mean approximates a normal distribution as the sample size increases no matter*

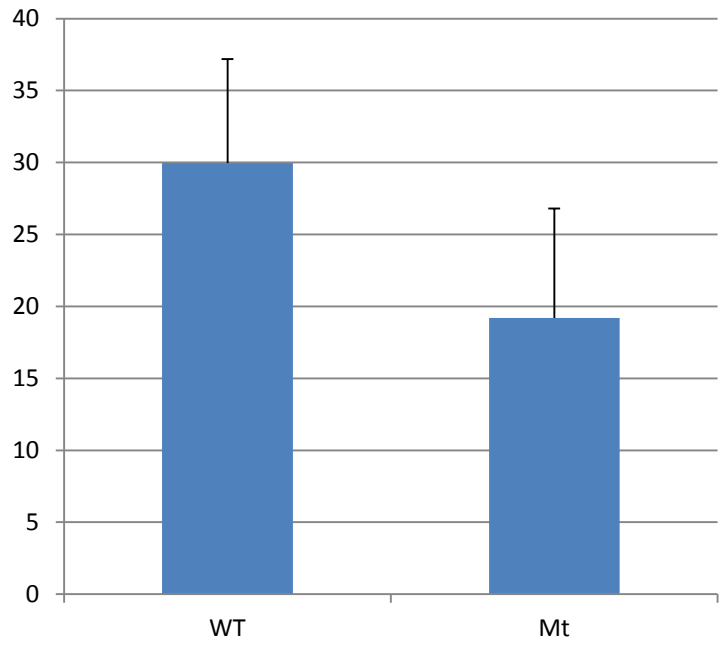
*2) Standard error of the mean decreases as the sample size increases:*



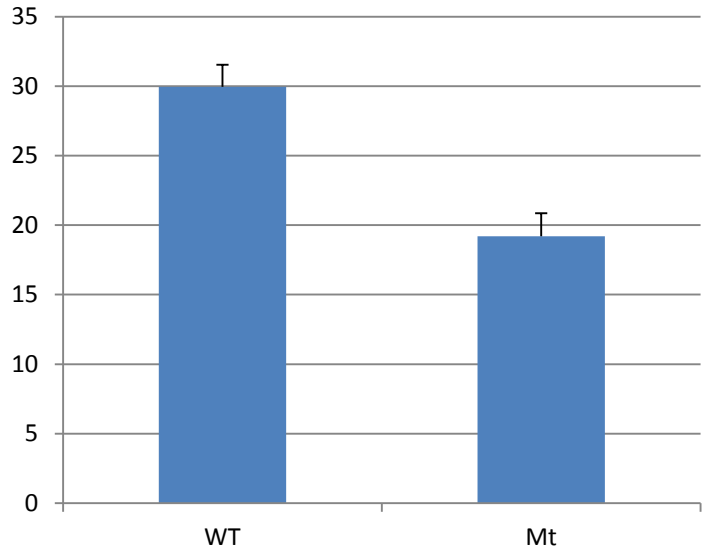
*Comparison between the standard deviation (SD) and the standard error of the mean (s.e.m):*

*The same data of WT and mt mice*

*Bar plot with **SD** as the error bars*




*Bar plot with **s.e.m** as the error bars*



# Poisson Application to Serial Dilution Assays

Got  a flat tire by a nail in Chicago last weekend.

What is the probability of such a flat tire in one block?

What is the probability of such a flat tire in five blocks?

Poisson Distribution

# Biological Examples of Poisson Distribution Application in

- The Number of  breast cancer stem cells in a  population
- The number of bacterial colonies in a Petri dish;
- The number of offspring an individual has;
- The number of nucleotide base substitutions in a gene over a period of time

Case:

How many stem cells in a breast cancer cell line culture?

The scientist made injections of the orignial cell Culture and 3 different dilutions into mice.

Serial Dilution Assay:

For each concentration of the cell culture, 8 mice were injected. The number of mice without tumor growth were Counted.

| dilution | #of mice out of 8 without tumor growth |
|---|---|
| original | 0 |
| 1/10 dilution | 3 |
| 1/100 dilution | 6 |
| 1/200 dilution | 7 |

| dilution | #of mice out of 8 without tumor growth | Observed Proportion |
|---|---|---|
| original | 0 | 0 |
| 1/10 dilution | 3 | 0.375 |
| 1/100 dilution | 6 | 0.75 |
| 1/200 dilution | 7 | 0.875 |

- Suppose $Y_1$ (the # of stem cell number in the original culture) follows a Poisson Distribution with mean=$u$

No event (No stem injected or no tumor growth) occurrence is equivalent to $Y_1 = 0$. The probability for $Y_1 = 0$ is

$$\Pr\{Y_1=0\}= e^{-u}$$

Another nice property of Poisson distribution

- For a dilution by t times (e.g., 10, 100, 200), the $Y_2$ (the # of stem cell number in the diluted culture) also follows a Poisson Distribution with mean=$u$/t

No event (No stem injected or no tumor growth) occurrence is equivalent to $Y_2 = 0$. The probability for $Y_2 = 0$ is

$$\Pr\{Y_2=0\}= e^{-u/t}$$

| dilution | #of mice out of 8 without tumor growth | Observed Proportion | Expected Proportion |
|---|---|---|---|
| original | 0 | 0 | |
| 1/10 dilution | 3 | 0.375 | exp(-u/10) |
| 1/100 dilution | 6 | 0.75 | exp(-u/100) |
| 1/200 dilution | 7 | 0.875 | exp(-u/200) |

Take the base e log of the observed and expected proportions

| dilution | #of mice out of 8 without tumor growth | Log Observed Proportion | Log Expected Proportion |
|---|---|---|---|
| original | 0 | 0 | |
| 1/10 dilution | 3 | $\text{Log}_e(0.375)$ | -u/10 |
| 1/100 dilution | 6 | $\text{Log}_e(0.75)$ | -u/100 |
| 1/200 dilution | 7 | $\text{Log}_e(0.875)$ | -u/200 |

Multiply through by the dilutions to get

| dilution | #of mice out of 8 without tumor growth | Log Observed Proportion | Log Expected Proportion |
|---|---|---|---|
| original | 0 | 0 | |
| 1/10 dilution | 3 | $-10*\text{Log}_e(0.375)$ | u |
| 1/100 dilution | 6 | $-100*\text{Log}_e(0.75)$ | u |
| 1/200 dilution | 7 | $-200*\text{Log}_e(0.875)$ | u |

| dilution | #of mice out of 8 without tumor growth | Log Observed Proportion | Log Expected Proportion |
|---|---|---|---|
| original | 0 | 0 | |
| 1/10 dilution | 3 | 9.8 | u |
| 1/100 dilution | 6 | 28.8 | u |
| 1/200 dilution | 7 | 26.7 | u |
| Mean | | 21.7 | u |

The true but unknown # of stem cells in the original culture is estimated to be 21.7.

# Does statistics help?

If Mr. Trump looked back at the polling stats of last Feb's Iowa Republican Caucus, what would he say?

"There are three types of lies -- lies, damn lies, and ***statistics***."

— Benjamin Disraeli