Summary Statistics & Sample Size Estimation

Panuwat Lertsithichai
Department of Surgery
Ramathibodi Hospital

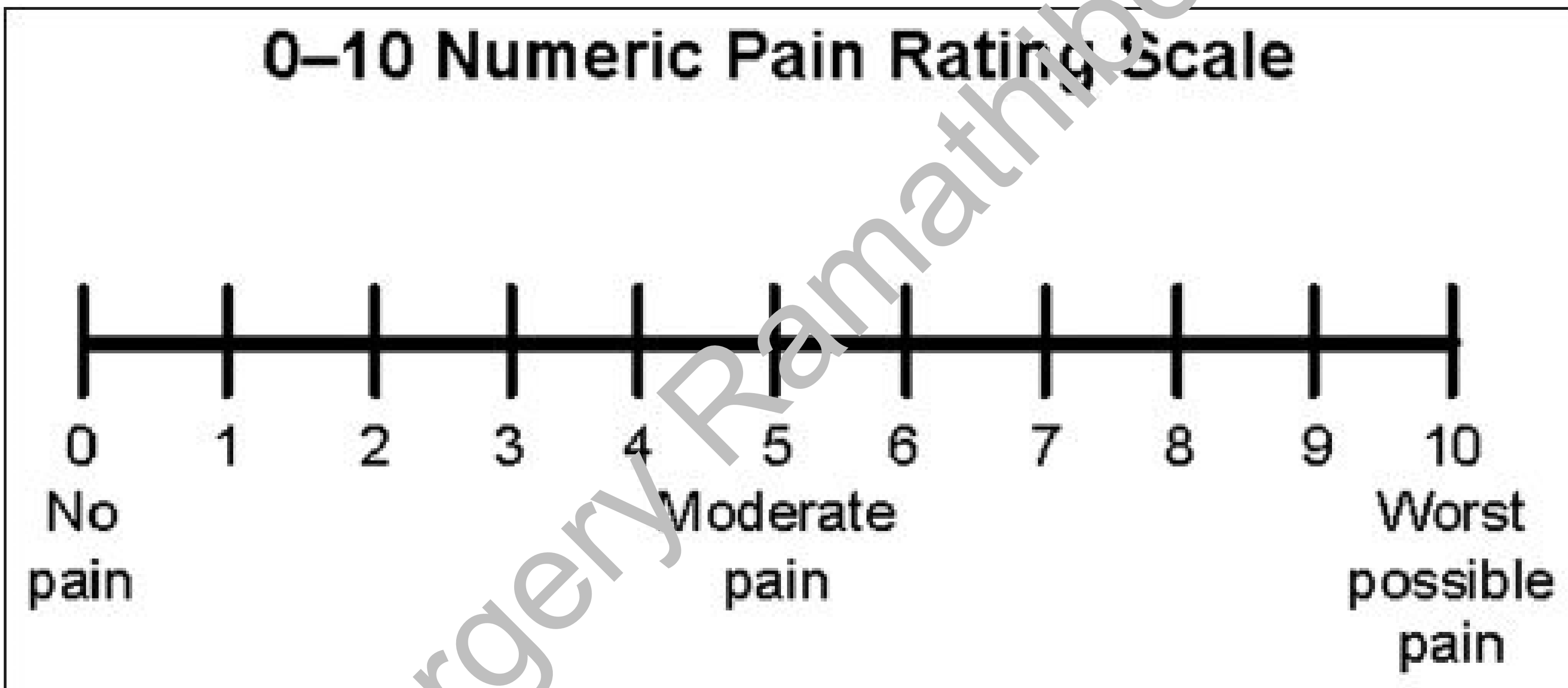
16 Sept 2020

Topics

- Summary statistics
- Statistical test
- Sample size estimation

A Clinical Trial

- A study comparing two analgesic drugs in terms of postoperative pain, A and B
- Designed as a parallel-group, randomized, controlled trial
- The pain is evaluated at 12 hours postoperative
- Using the Visual Analog Scale (VAS)

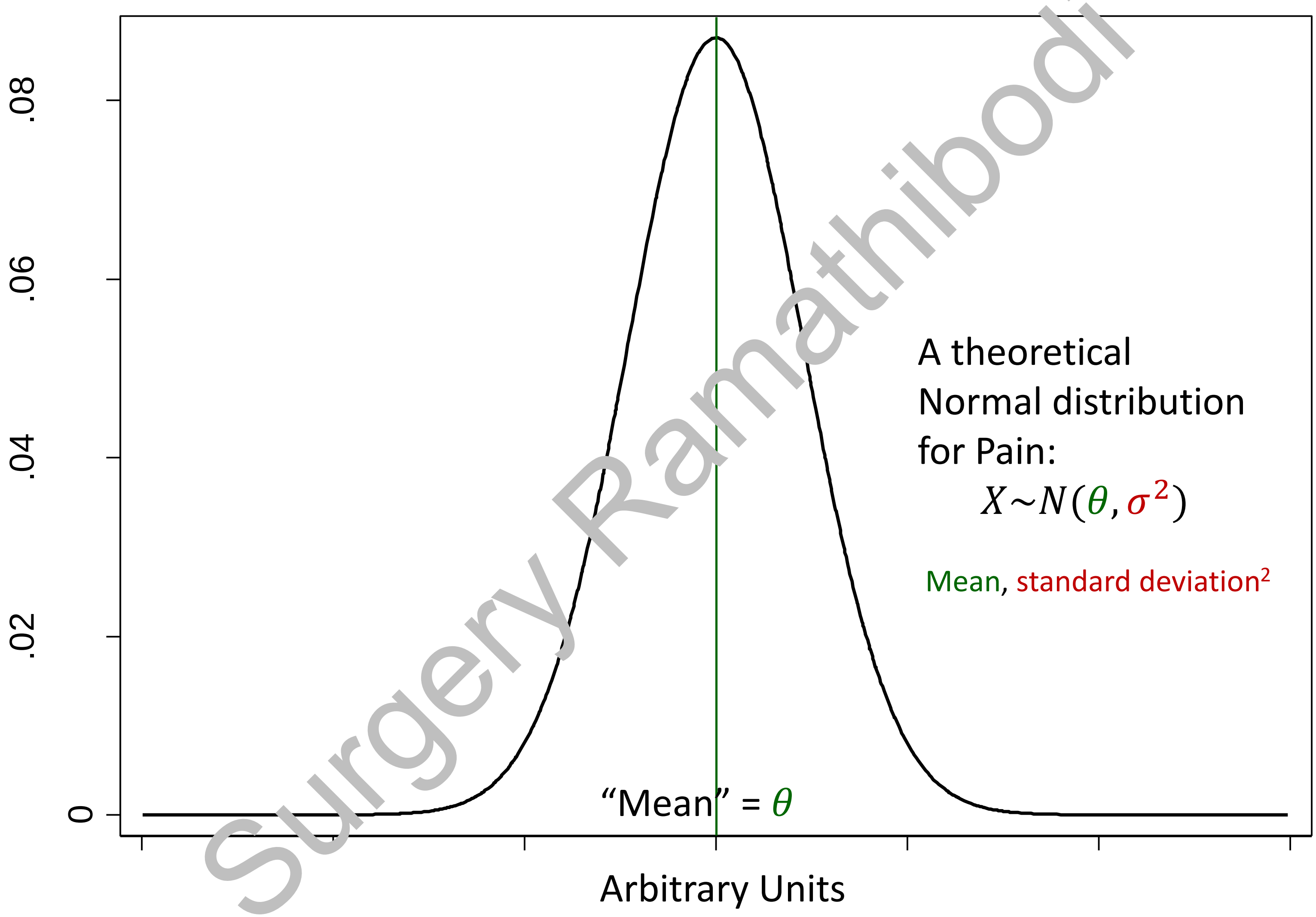

0–10 Numeric Pain Rating Scale
0
1
2
3
4
5
6
7
8
9
10
No pain
Moderate pain
Worst possible pain

Outcome Variable: Pain

- Pain (VAS) can be considered a quantitative variable
- Measurements on *different* patients undergoing the same operation will have *different* values
- These values form a **distribution**
- *In theory*, the possible values are infinite, and may often have an approximately Normal distribution*

*We follow the Frequentist Parametric (model-based) Approach today

“Normal” (Gaussian) **Distribution**



Intervention (Exposure) Variable

- Analgesic drug: A; B
- Random assignment (simple randomization)

Patient Characteristics/Baseline Variables

- Age: years
- Sex/gender: female; male (0,1)
- Extent/type of operation: small; big (0,1)
- Preoperative anxiety: no; yes (0,1)

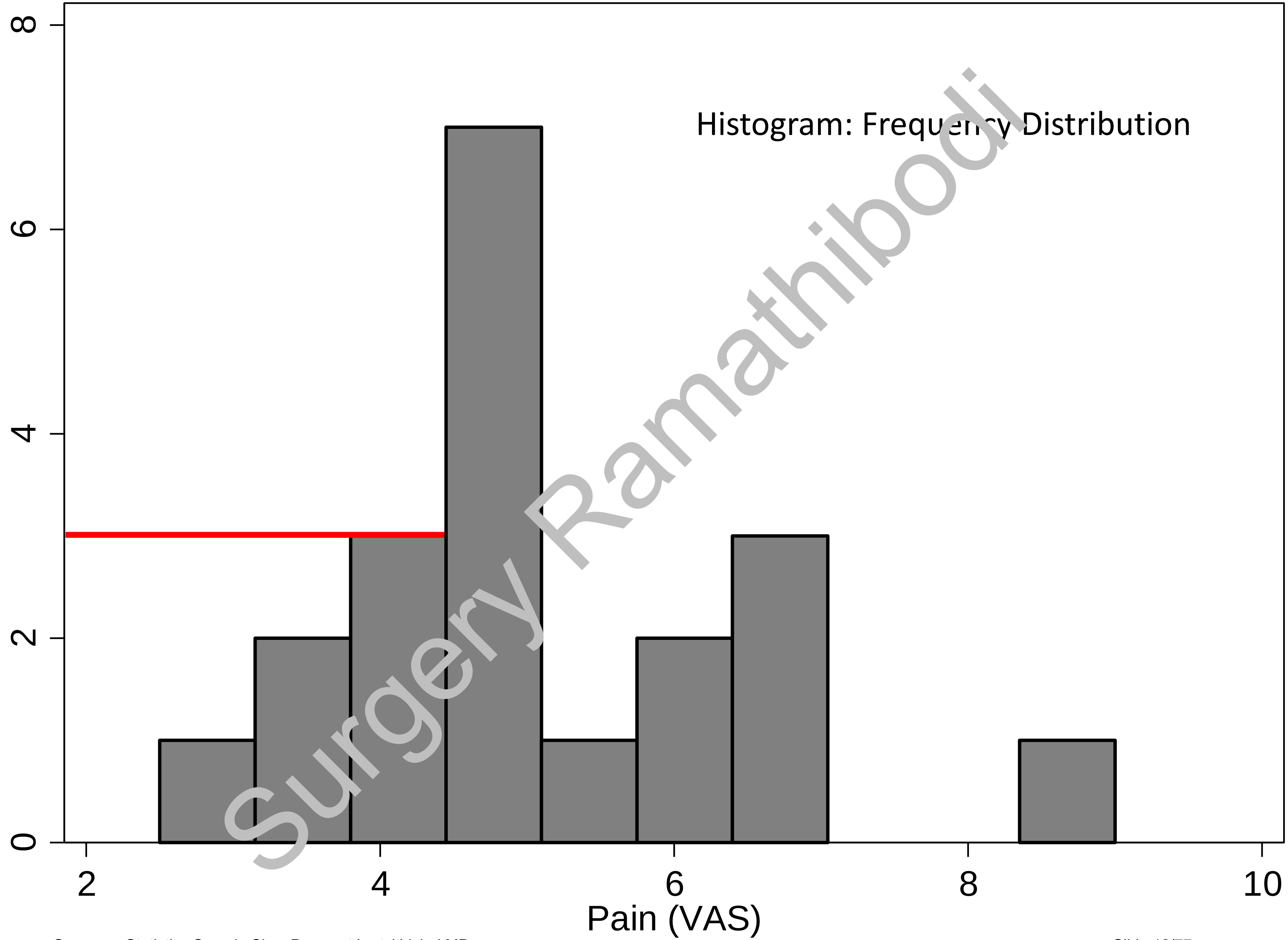
	id	age	sex	anxiety	operation	Drug	Pain
1	1	48	0	1	0	A	4.5
2	2	46	1	0	0	B	5
3	3	43	1	1	0	A	6
4	4	39	1	1	0	B	7
5	5	57	0	0	0	B	4
6	6	48	0	1	0	B	5
7	7	41	1	0	0	B	2.5
8	8	47	0	0	0	A	3.5
9	9	51	1	1	1	A	9
10	10	57	0	0	1	A	4.5
11	11	45	0	1	0	A	5
12	12	57	1	0	1	B	6
13	13	51	1	0	1	B	7
14	14	48	0	0	1	A	7
15	15	56	0	0	0	A	5
16	16	48	1	0	0	B	3.5
17	17	4[illegible]	1	1	0	A	5.5
18	18	54	0	0	0	B	4
19	19	53	0	0	0	A	4
20	20	53	0	0	0	B	4.5

	id	age	sex	anxiety	operation	Drug	Pain
1	1	48	0	1	0	A	4.5
2	2	46	1	0	0	B	5
3	3	43	1	1	0	A	6
4	4	39	1	1	0	B	7
5	5	57	0	0	0	B	4
6	6	48	0	1	0	B	5
7	7	41	1	0	0	B	2.5
8	8	47	0	0	0	A	3.5
9	9	51	1	1	1	A	9
10	10	57	0	0	1	A	4.5
11	11	45	0	1	0	A	5
12	12	57	1	0	1	B	6
13	13	51	1	0	1	B	7
14	14	48	0	0	1	A	7
15	15	56	0	0	0	A	5
16	16	48	1	0	0	B	3.5
17	17	42	1	1	0	A	5.5
18	18	54	0	0	0	B	4
19	19	53	0	0	0	A	4
20	20	53	0	0	0	B	4.5

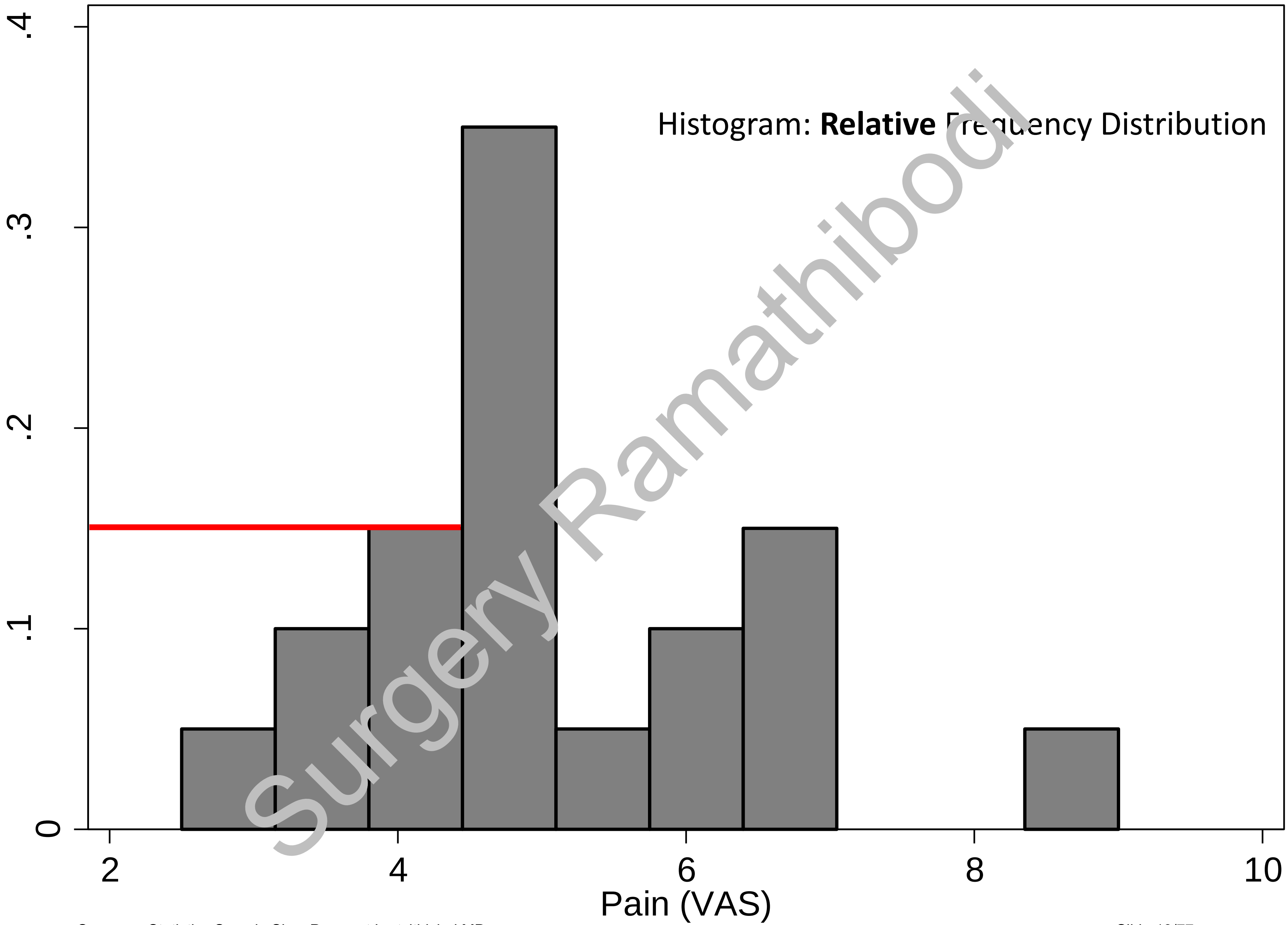
Summarizing & Displaying Variables

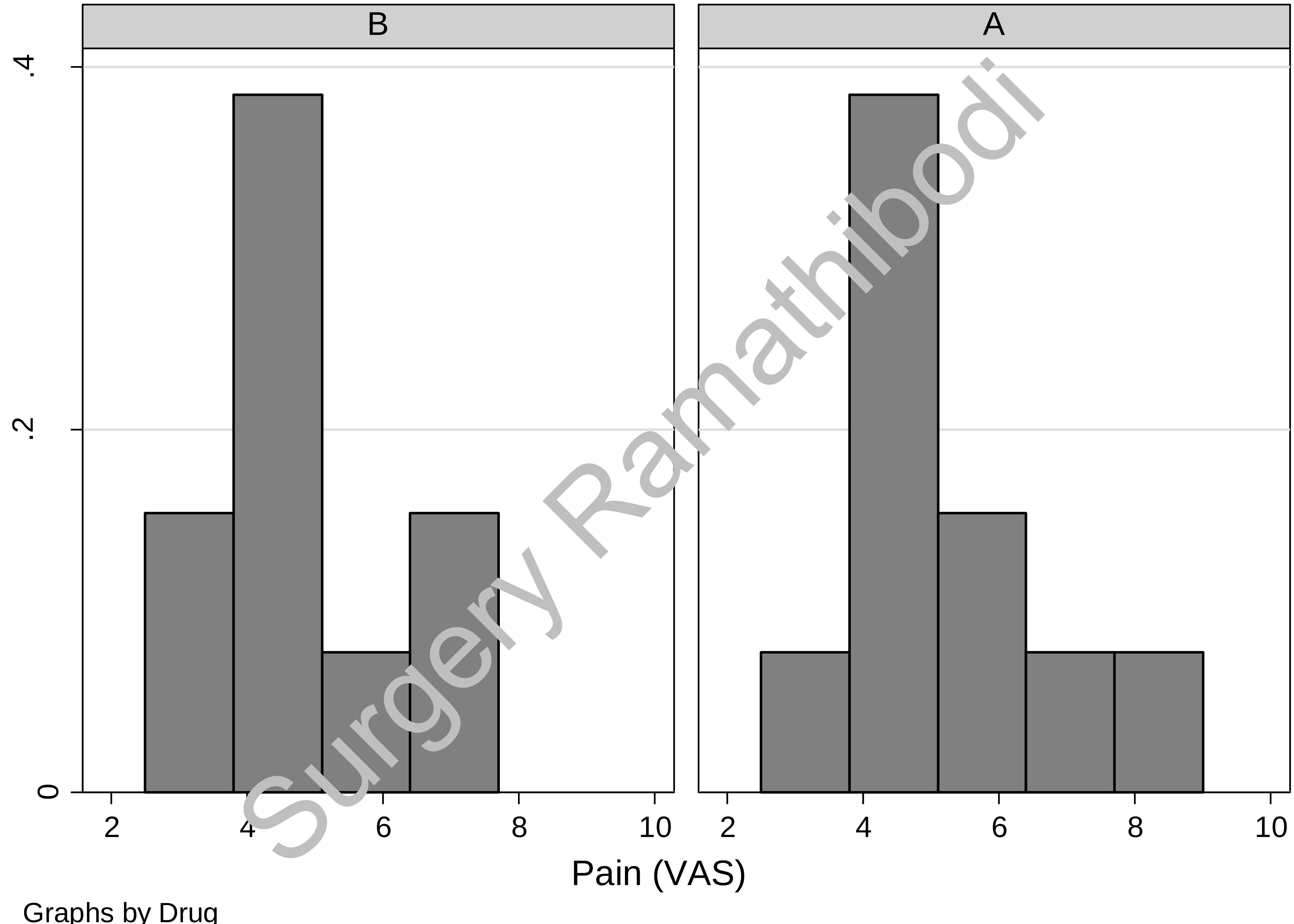
- **Statistics** (Numbers)
- **Graphics**: Histograms; box-whisker plots; stem-leaf; scatter plots; line graphs; etc.

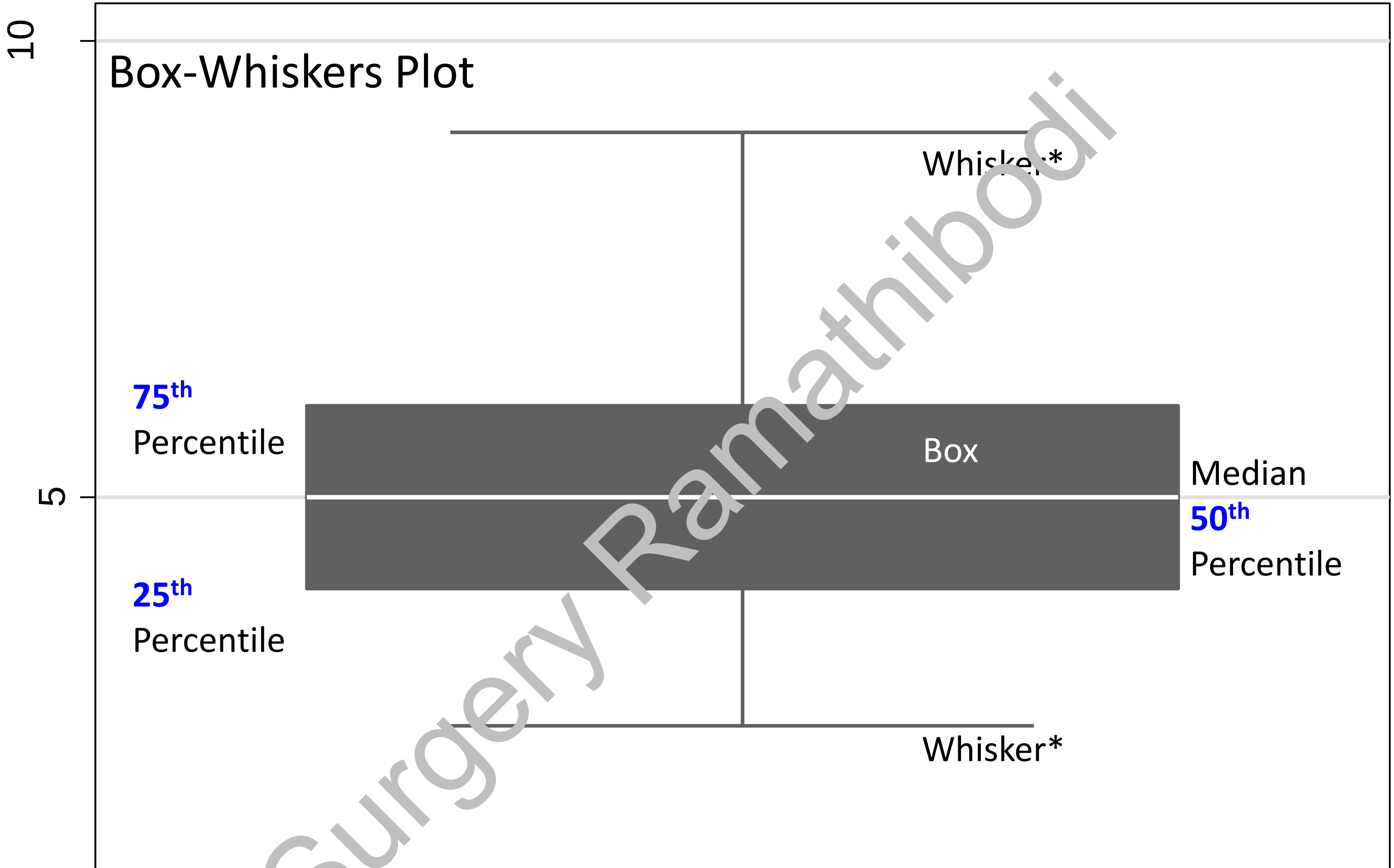
Choice of most appropriate summary or display depends on type of data & study objectives



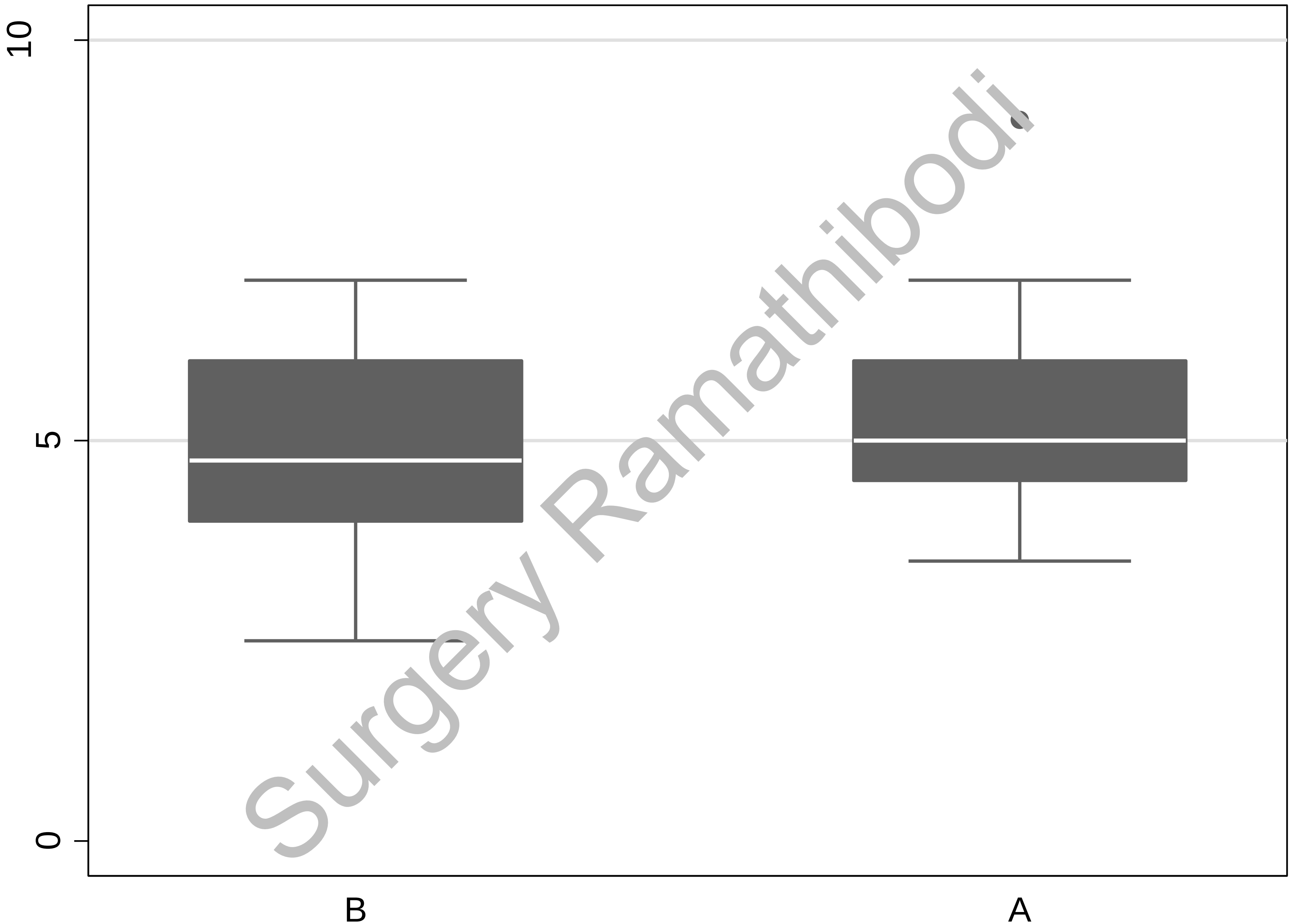
Histogram: Frequency Distribution
0
2
4
6
8
2
4
6
8
10
Pain (VAS)


Histogram: **Relative** Frequency Distribution
0
.1
.2
.3
.4
2
4
6
8
10
Pain (VAS)


B
A
.4
.2
0
2
4
6
8
10
2
4
6
8
10
Pain (VAS)
Graphs by Drug



*Whiskers are capped above at the value x_u where $x_u = x_{(i)} \leq x_{(75)} + 1.5 \times \mathrm{IQR}$ with $x_{(i+1)} > x_{(75)} + 1.5 \times \mathrm{IQR}$;
Whiskers are capped below at the value x_l where $x_l = x_{(i)} \geq x_{(25)} - 1.5 \times \mathrm{IQR}$ with $x_{(i-1)} < x_{(25)} - 1.5 \times \mathrm{IQR}$;
Where (i) represents i^{th} ordered value; (25) represents the 25^{th} percentile value, etc. (this is from John W. Tukey)


10
5
0
B
A

Stem-Leaf Plots

Overall

```
2* | 5
3* | 55
4* | 000555
5* | 00005
6* | 00
7* | 000
8* |
9* | 0
```

B

```
2* | 5
3* | 5
4* | 005
5* | 00
6* | 0
7* | 00
```

A

```
3* | 5
4* | 055
5* | 005
6* | 0
7* | 0
8* |
9* | 0
```

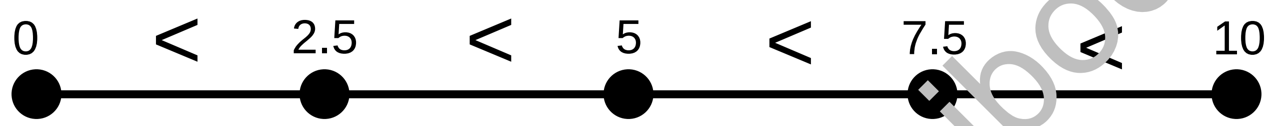
Why Summary Statistics

- Better for interpreting, and comparing between, **groups**
- Can be used as a basis for *statistical tests* and *sample size calculations*

Classification of Variables

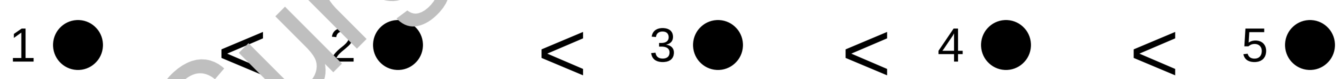
- “Randomness”: Random; non-random variables
- Research-related: Outcome; exposure; baseline
- “Arithmetical”: Quantitative; categorical; ordinal

Arithmetical Type of Variables



Quantitative data: Pain VAS – *continuous*

Nominal Categorical Data: Occupation



Ordinal Data: Pain Level

Outcome Variable(s)

- Pain score: Visual Analogue Scale (from 1 to 10)
- Pain level: 2 levels; not very painful (0); very painful (1)

  : 5 levels; 1, 2, 3, 4, 5

Exposure Variable

- Analgesic drug: A(1); B(2)

Baseline Variables

- Age: years (quantitative, continuous variable)
- Sex/gender: 0, 1
- Extent/type of operation: 0, 1
- Preoperative anxiety: 0, 1

Exercise

Which arithmetical type of variable ?

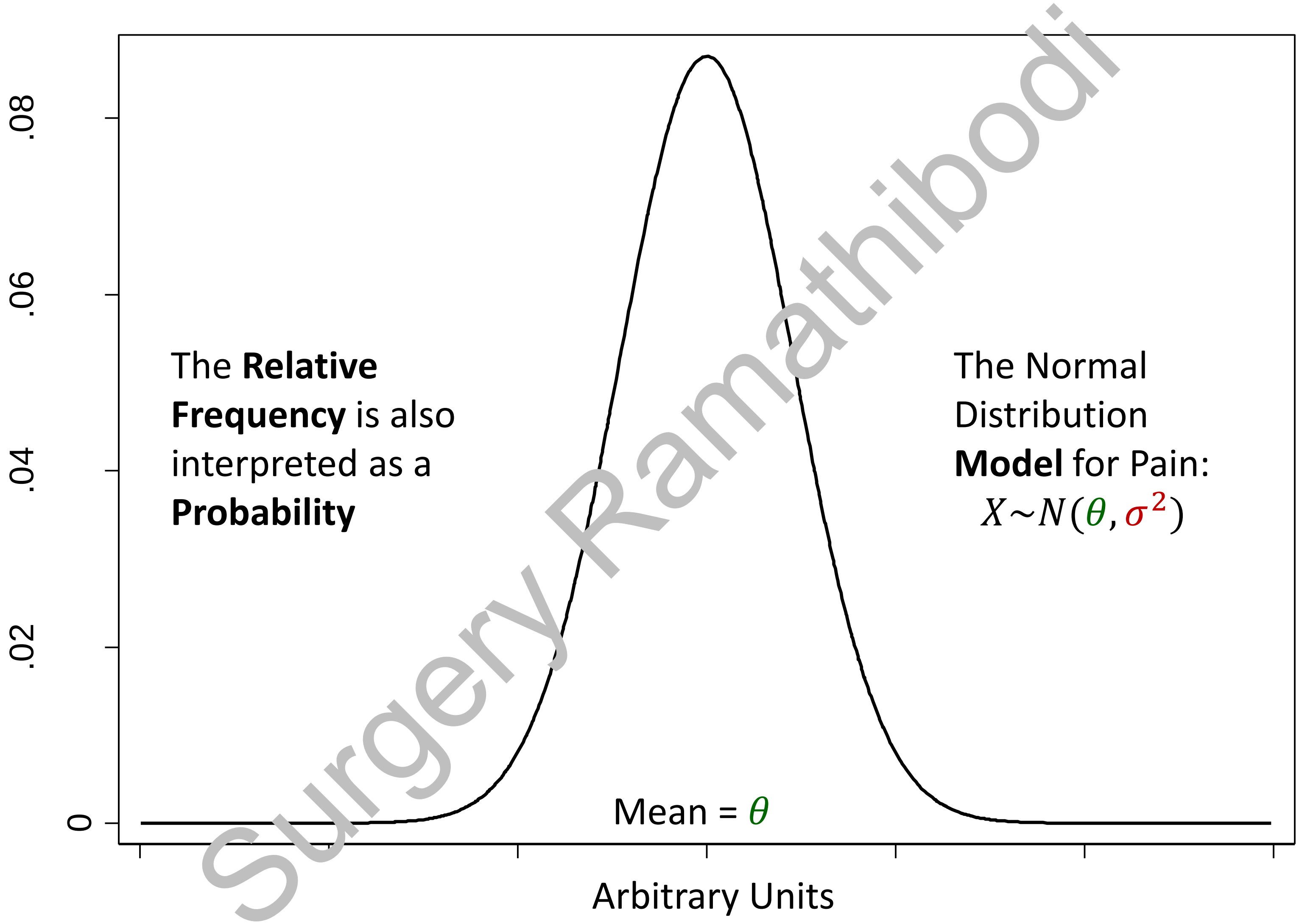
- Height
- Age categorized as 5-10, 11-20, 21-30, etc.
- Number of people in any district
- Types of University Education in a sample
- Complications after a surgical procedure
- Socioeconomic status of a group of people

Summarizing Quantitative Variables

At least two summary measures:

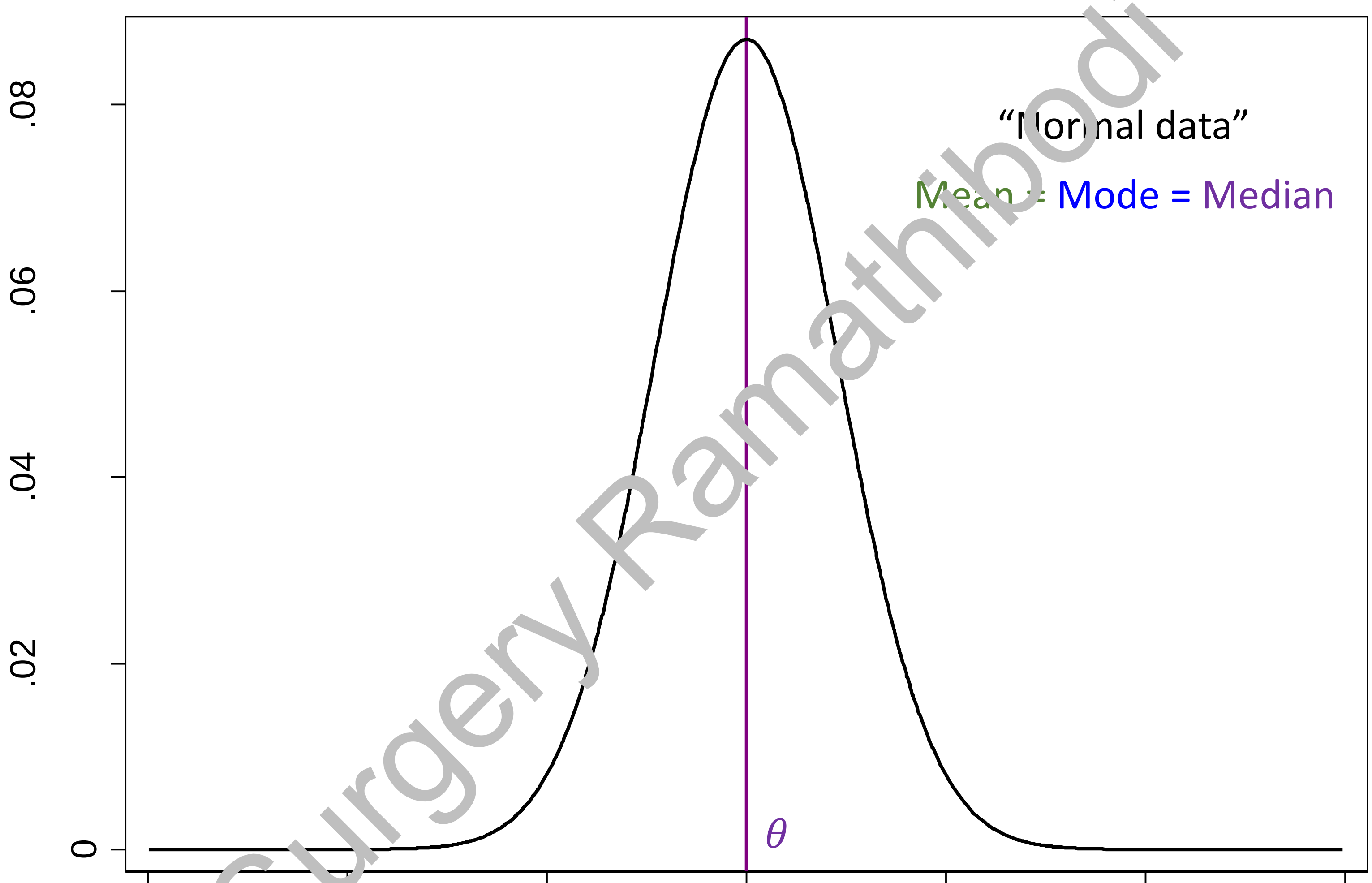
- A measure of “central tendency”
  - **Mean; median**
- A measure of “spread”/data variation
  - **Standard deviation (sd); range**

“Normal” (Gaussian) **Relative Frequency Distribution (Probability Density*)**

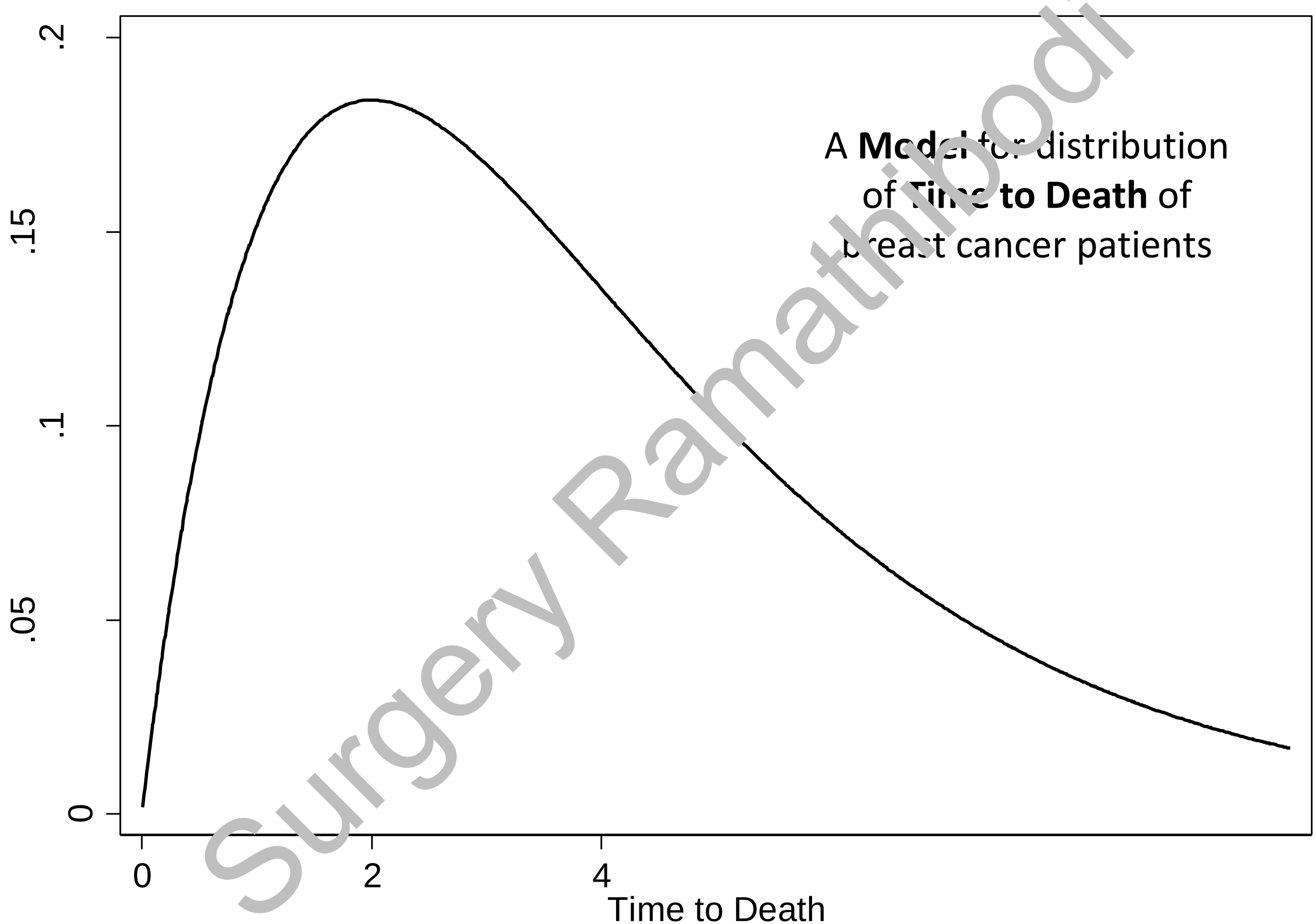


*The relative frequency is substituted by the density in the continuous-value limit

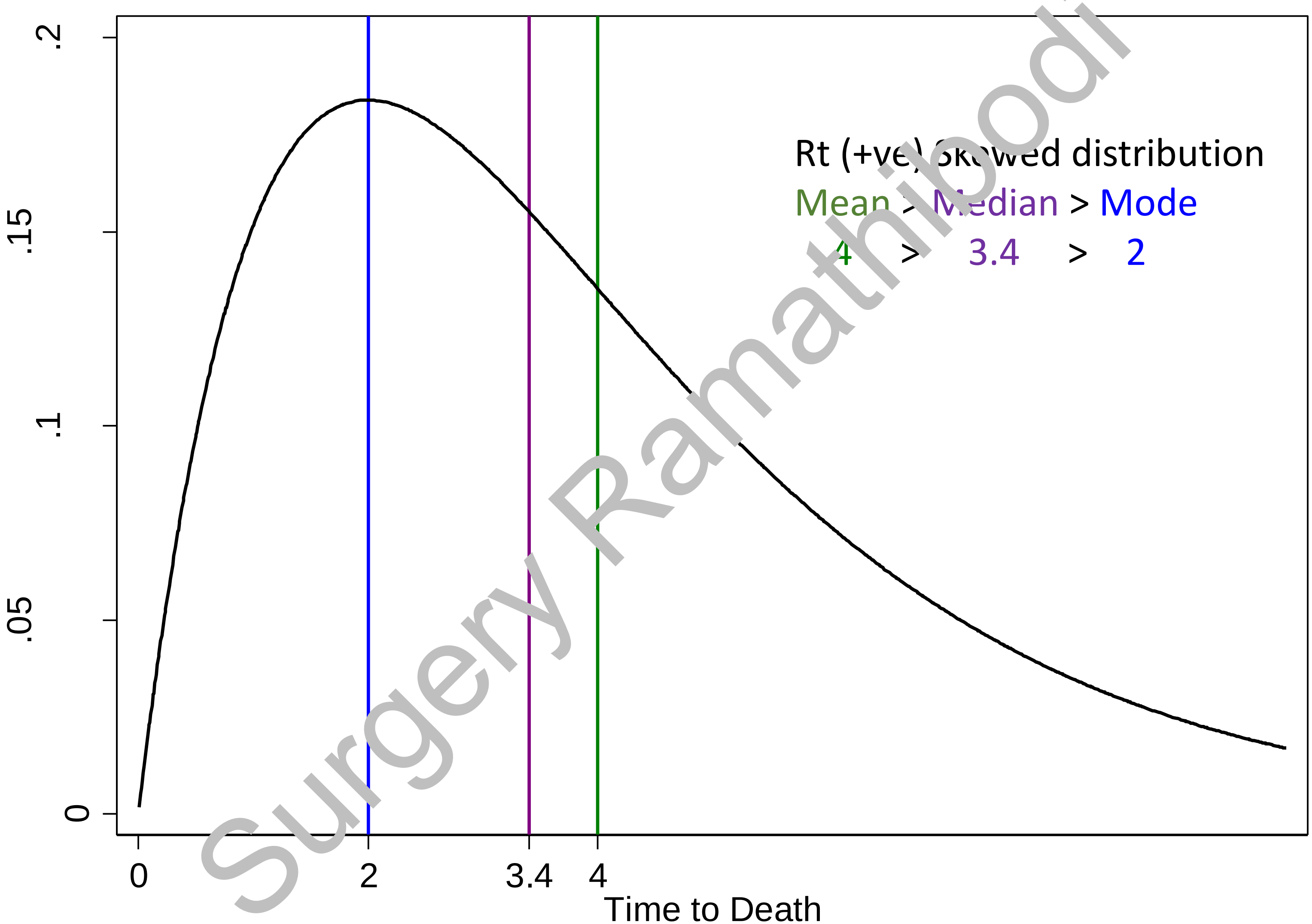
“Central Tendency”



Non-Normal Distribution



"Central Tendency"



Summarizing Pain (VAS)

Drug A: 3.5, 4, 4.5, 4.5, 5, 5, 5.5, 6, 7, 9

Drug B: 2.5, 3.5, 4, 4, 4.5, 5, 5, 6, 7, 7

n = 10/group

Mean; sd A: 5.4; 1.6 B: 4.9; 1.5

Med; range A: 5; [3.5 , 9] B: 4.75; [2.5, 7]

IQR* A: [4.5, 6] B:[4, 6]

*Interquartile Range

Time to Event (Death)

Operation A (months)

- 1, 2, 2, 3, 4, 7, 17, 20, 45 ; n = 9 ; mean = 11.2;
- Median = 4

Operation B (months)

- 1, 2, 3, 5, 7, 12, 18, 21, 27 ; n = 9; mean = 10.7;
- Median = 7

Which summary is more appropriate?

Use of Median / Range

Possible rationale

- **Consistency**: if *Rank tests* (Non-parametric tests) are used, then summary statistics should be based on Ranks (Order statistics)
- **Robustness** to "outliers"

2, 5, 9, 17, **29**, 91, 180, 392, 901 ; mean = **180.7**

2, 5, 9, 17, **29**, 91, 180, 392, 9801 ; mean = **1169.6**

Normal or Non-Normal ?

The use of mean & sd are appropriate when

1. **Mean is at least 2×sd**
2. **Mean ~ median**
3. Histogram is Normal-shaped
4. Variable can assume positive and negative values

Example

Median ~ Mean and Mean > 2×sd ?

Age (years):

Mean; sd	A: 49.1; 5.0	B: 49.4; 6.2
Median	A: 48	B: 49.5

Serum albumin (gm/dL):

Mean; sd	5.2; 6.7
Median	3.5

Exercise

Summarize these quantitative variables

32, 34, 39, 42, 45, 47, 53, 58, 60, 60, 61, 63

0,0,0,0,0,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,4,4

68, 79, 80, 88, 98, 110, 134, 155, 160, 230, 347

Testing Quantitative Variables

- Use **standardized mean difference** of quantitative, "Normal", variables to test for significant difference between groups

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(sd_A^2 + sd_B^2\right)/n}} \equiv t$$

Mean for group A

The ***t*-statistic**

Standard deviation for group A

Sample size per group

Testing Quantitative Variables

- Use **standardized mean difference** of quantitative, "Normal", variables to test for significant difference between groups

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(sd_A^2 + sd_B^2\right)/n}} \equiv t$$

Mean difference

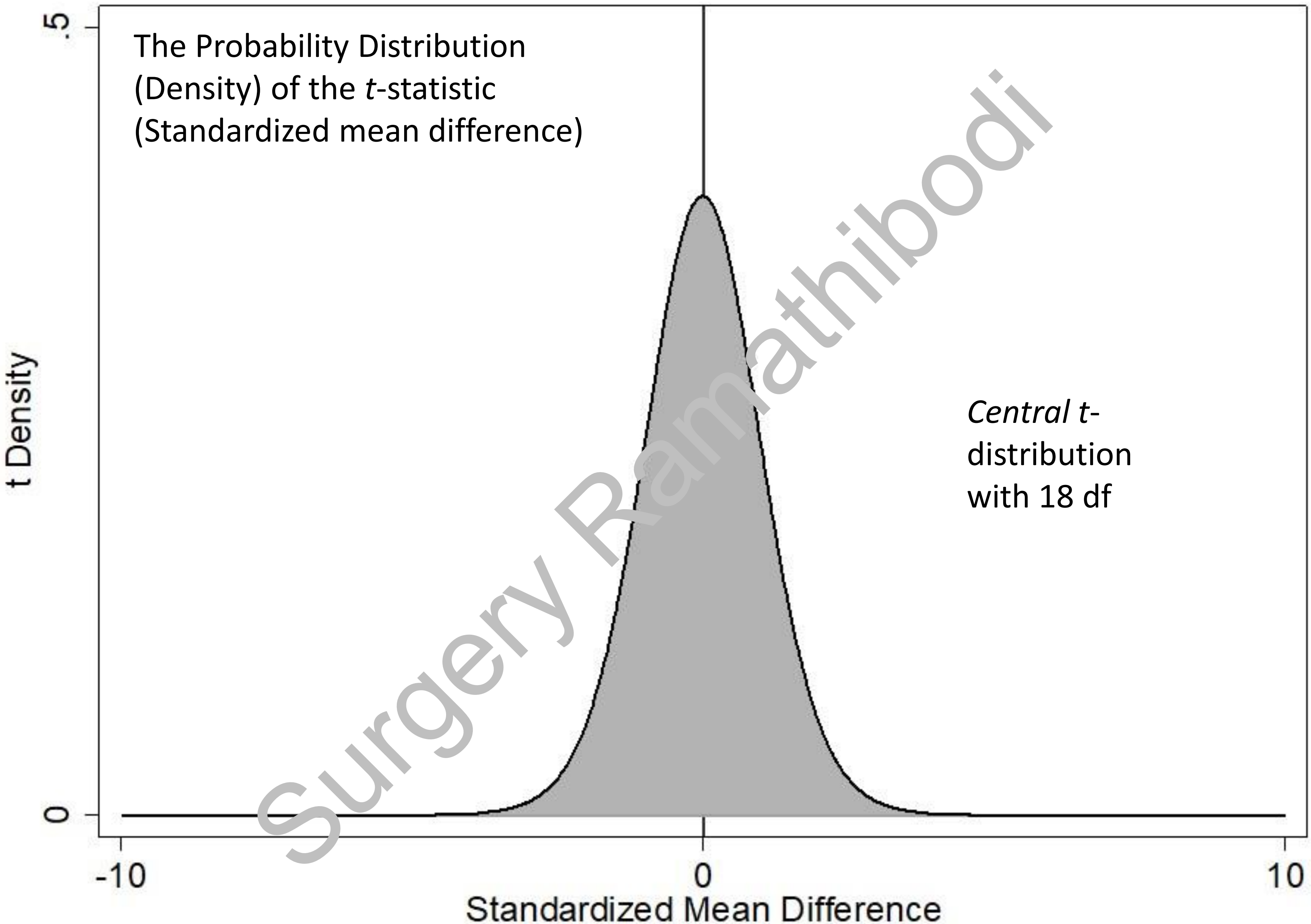
Standard error of the mean difference

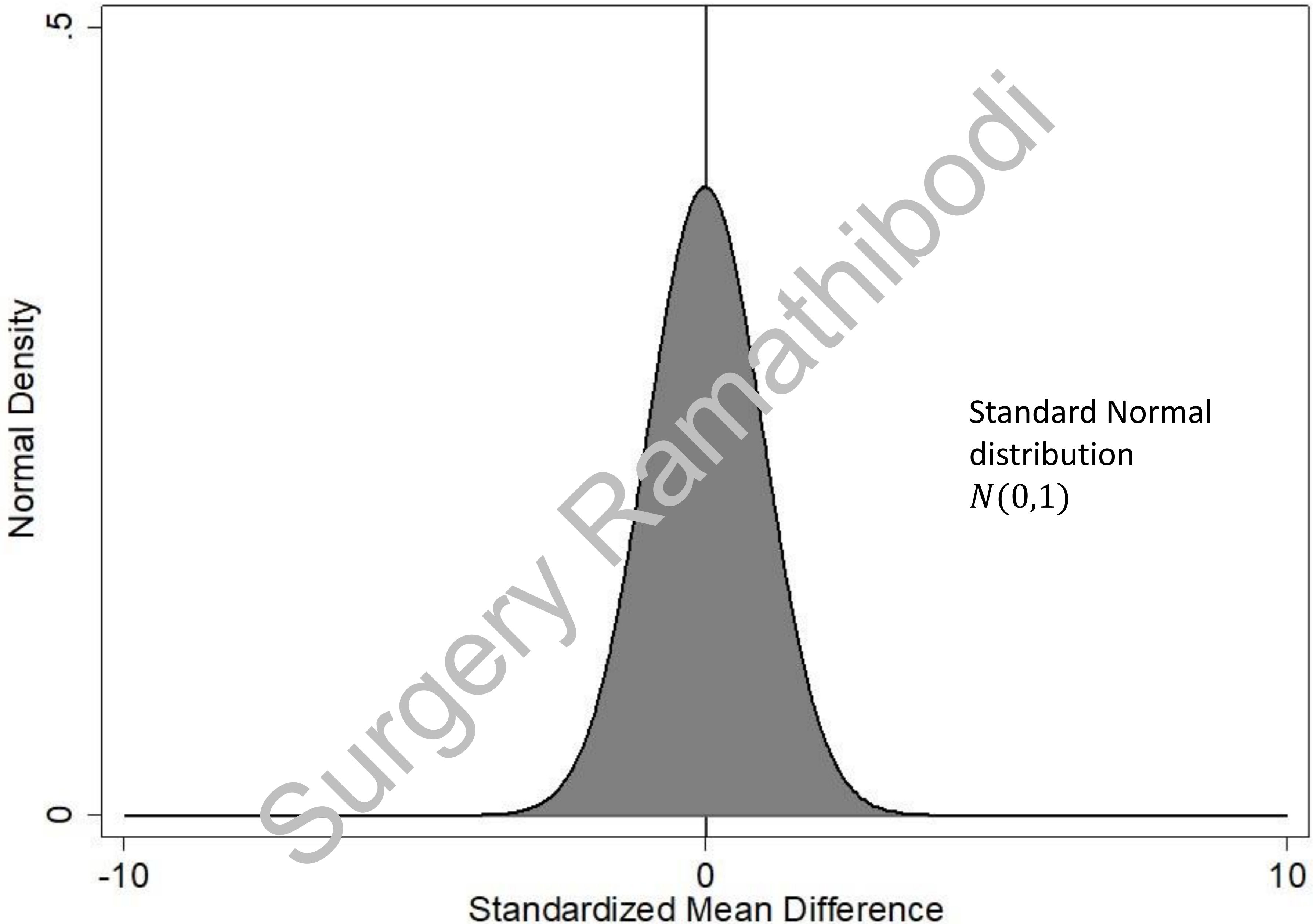
Testing Quantitative Variables

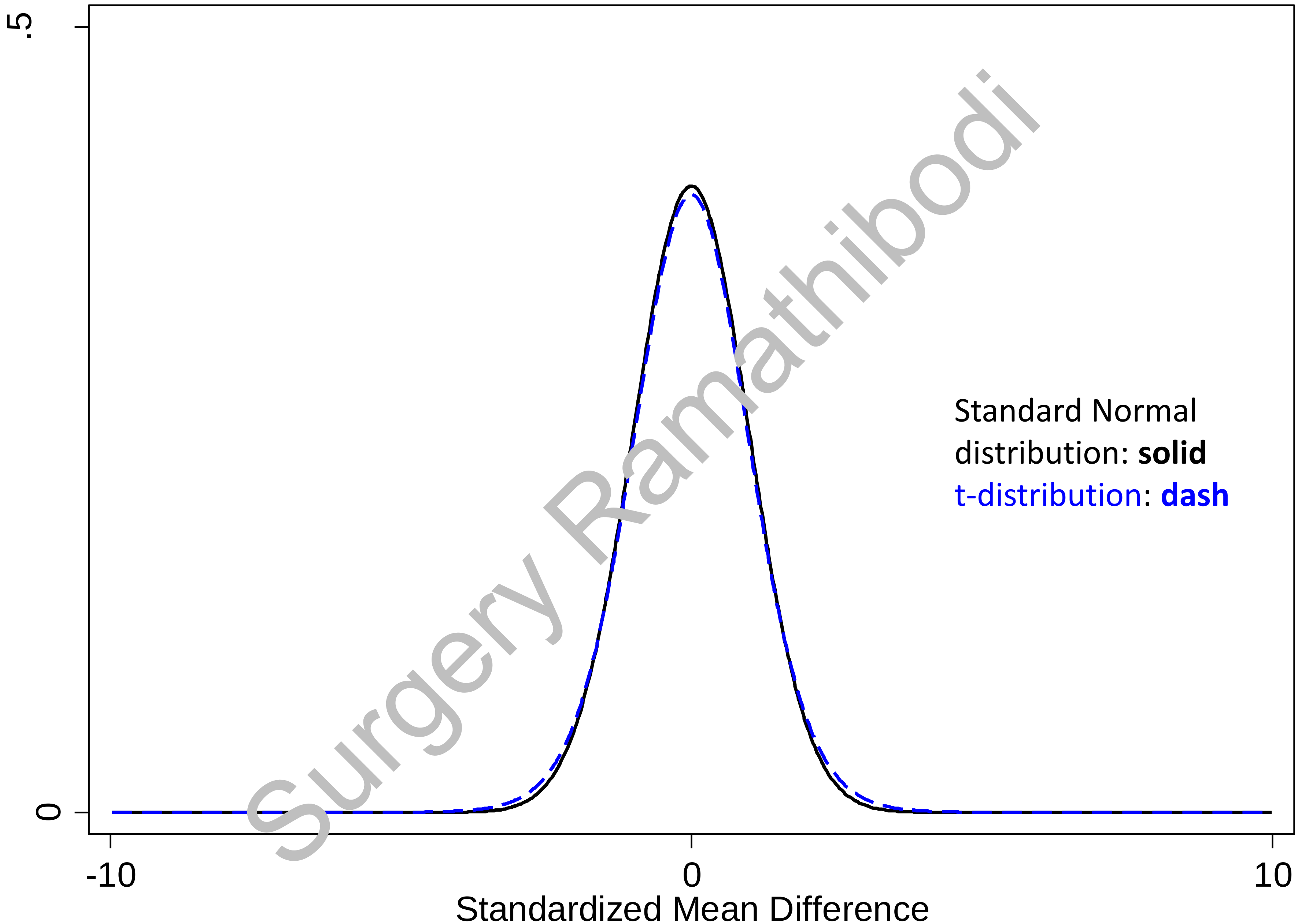
- Use **standardized mean difference** of quantitative, "Normal", variables to test for significant difference between groups

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(sd_A^2 + sd_B^2\right)/n}} \equiv t$$

- The standardized mean difference, viewed as a *random variable*, has a *t*-distribution ($2n - 2$ df)
- If we hypothetically perform identical experiments many, many times the values of the *t*-statistic will form a probability distribution


The Probability Distribution
(Density) of the *t*-statistic
(Standardized mean difference)
Central t-
distribution
with 18 df
t Density
.5
0
-10
0
10
Standardized Mean Difference


.5
0
Normal Density
-10
0
10
Standardized Mean Difference
Standard Normal
distribution
$N(0,1)$



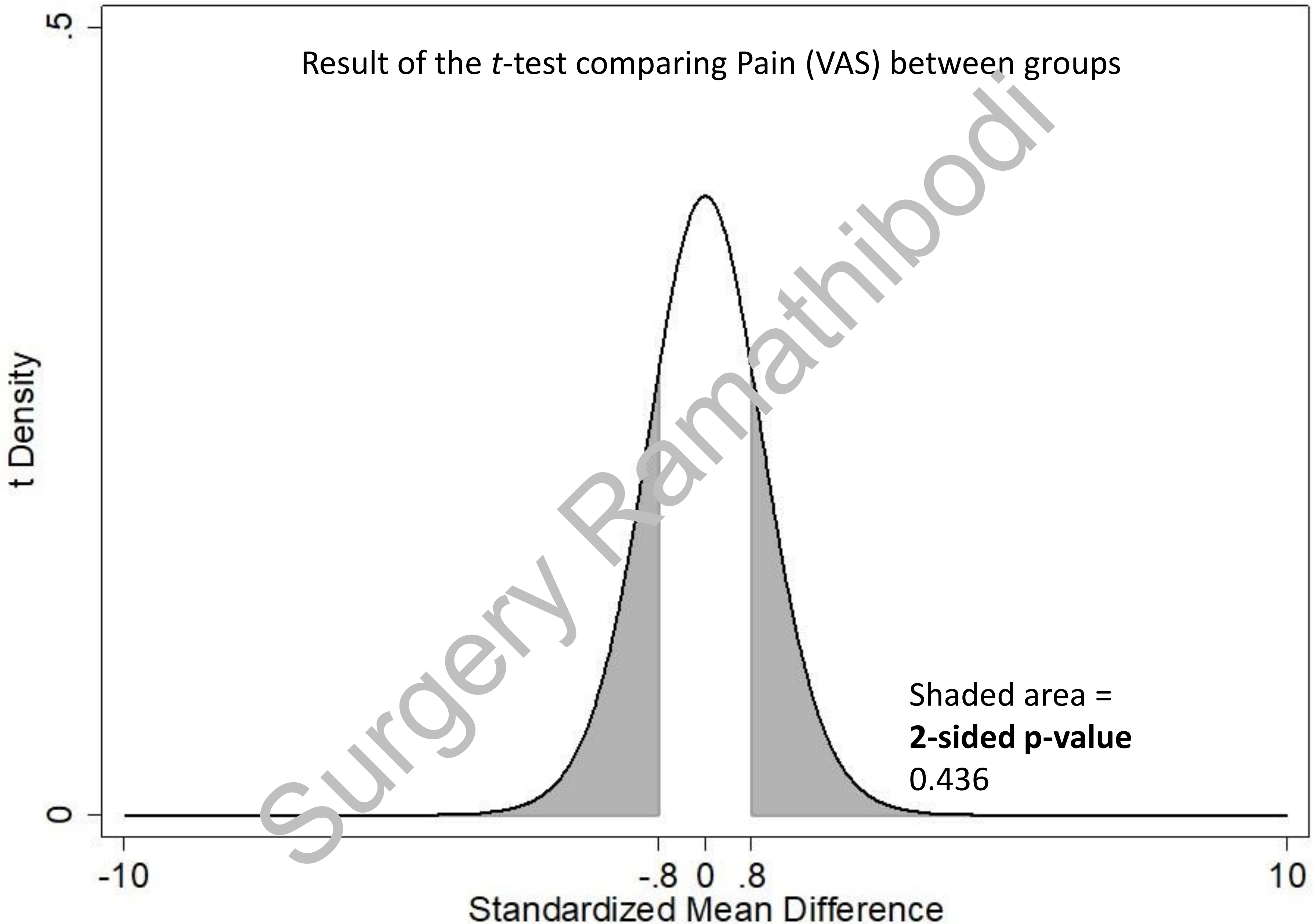
.5
0
-10
0
10
Standardized Mean Difference
Standard Normal distribution: **solid**
t-distribution: **dash**

Testing Pain (VAS)

- The *mean difference* is 5.4 – 4.9 = 0.5
- The *standardized mean difference* for **Pain (VAS)** is

$$\frac{5.4-4.9}{\sqrt{(1.5^2+1.6^2)/10}} = 0.80$$

Compare this to a central *t*-distribution (18 df); 2-sided **p-value = 0.436**



Result of the *t*-test comparing Pain (VAS) between groups
Shaded area =
2-sided p-value
0.436
.5
0
t Density
-10
-.8 0 .8
10
Standardized Mean Difference

	id	age	sex	anxiety	operation	Drug	Pain
1	1	48	0	1	0	A	4.5
2	2	46	1	0	0	B	5
3	3	43	1	1	0	A	6
4	4	39	1	1	0	B	7
5	5	57	0	0	0	B	4
6	6	48	0	1	0	B	5
7	7	41	1	0	0	B	2.5
8	8	47	0	0	0	A	3.5
9	9	51	1	1	1	A	9
10	10	57	0	0	1	A	4.5
11	11	45	0	1	0	A	5
12	12	57	1	0	1	B	6
13	13	51	1	0	1	B	7
14	14	48	0	0	1	A	7
15	15	56	0	0	0	A	5
16	16	48	1	0	0	B	3.5
17	17	41	1	1	0	A	5.5
18	18	54	0	0	0	B	4
19	19	53	0	0	0	A	4
20	20	53	0	0	0	B	4.5

Testing Age

- The *standardized mean difference* for **Age (years)** is

$$\frac{49.1-49.4}{\sqrt{(5.0^2+6.2^2)/10}} = -0.12$$

Compare this to a central *t*-distribution (18 df); 2-sided **p-value = 0.907**

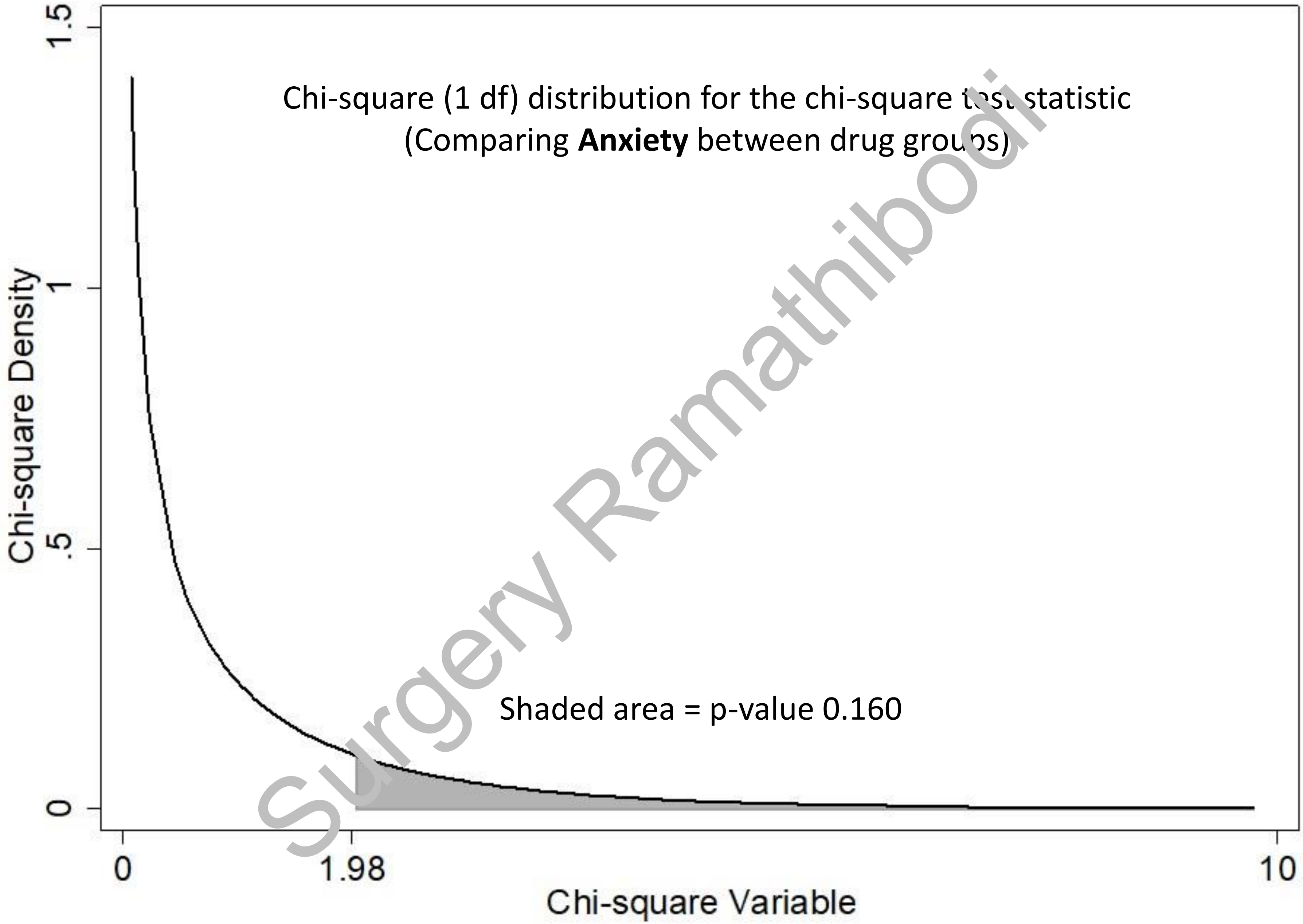
Summarizing Categorical Variables

Variable	Drug A	Drug B	p-value
Sex: Female Male	3 (30%) 7 (70%)	6 (60%) 4 (40%)	
Op: Large Small	3 (30%) 7 (70%)	2 (20%) 8 (80%)	
Anxiety: Yes No	5 (50%) 5 (50%)	2 (20%) 8 (80%)	

Testing Categorical Variables

Variable	Drug A	Drug B	p-value
Sex: Female Male	3 (30%) 7 (70%)	6 (60%) 4 (40%)	0.178* 0.370**
Op: Large Small	3 (30%) 7 (70%)	2 (20%) 8 (80%)	0.606* 0.999**
Anxiety: Yes No	5 (50%) 5 (50%)	2 (20%) 8 (80%)	0.160* 0.350**

* Chi-square test; ** Fisher's exact test


Chi-square (1 df) distribution for the chi-square test statistic
(Comparing **Anxiety** between drug groups)
Shaded area = p-value 0.160
Chi-square Density
1.5
1
.5
0
0
1.98
10
Chi-square Variable

Ordinal Variables

- Variables with $\leq$ 5 categories, use counts to summarize data*
- Use median & range for > 5 categories*
- When testing for differences, *efficiency* (i.e., greater *sensitivity* to detect differences if they exist) requires the use of ranks

*These are recommendations

Summarizing Ordinal Variables

Pain level	Drug B: n(%)	Drug A: n(%)
1	5 (22)	2 (9)
2	9 (39)	3 (13)
3	3 (13)	5 (22)
4	4 (17)	6 (26)
5	2 (9)	7 (30)
Total	23 (100)	23 (100)

Suppose we perform another study with 46 subjects, 23 per group, and compare the **Pain Level** between the two drugs

Testing Ordinal Variables

Pain level	Drug B: n(%)	Drug A: n(%)
1	5 (22)	2 (9)
2	9 (39)	3 (13)
3	3 (13)	5 (22)
4	4 (17)	6 (26)
5	2 (9)	7 (30)
Total	23 (100)	23 (100)

Chi-square test p-value = 0.093

Fisher's test p-value = 0.099

Ranksum test p-value = 0.011

t-test p-value = 0.009

What's The Interpretation?

Let's come back to the results of the present study:

- No significant differences in baseline characteristics
- No significant difference in Pain (VAS)

The *Observed* Mean Difference

- The ***observed*** mean difference is $\bar{x}_A - \bar{x}_B = 0.5$
- The ***observed standardized mean difference*** is not "statistically significant":

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(sd_A^2 + sd_B^2\right)/n}} = 0.8$$

- The sample size, $n = 10$/gr, is probably too small
- Then how large should a sample size be?

The Sample Size

- An appropriate **sample size** (from a *Frequentist* perspective) *is such that*:
- If a true difference exists, the experiment is able to detect a significant difference (at 5% level) at least 80% of the time (80% sensitivity)
- The 5% and 80% are conventional numbers representing acceptable errors (type I error = 5% & type II error = 20%) in research

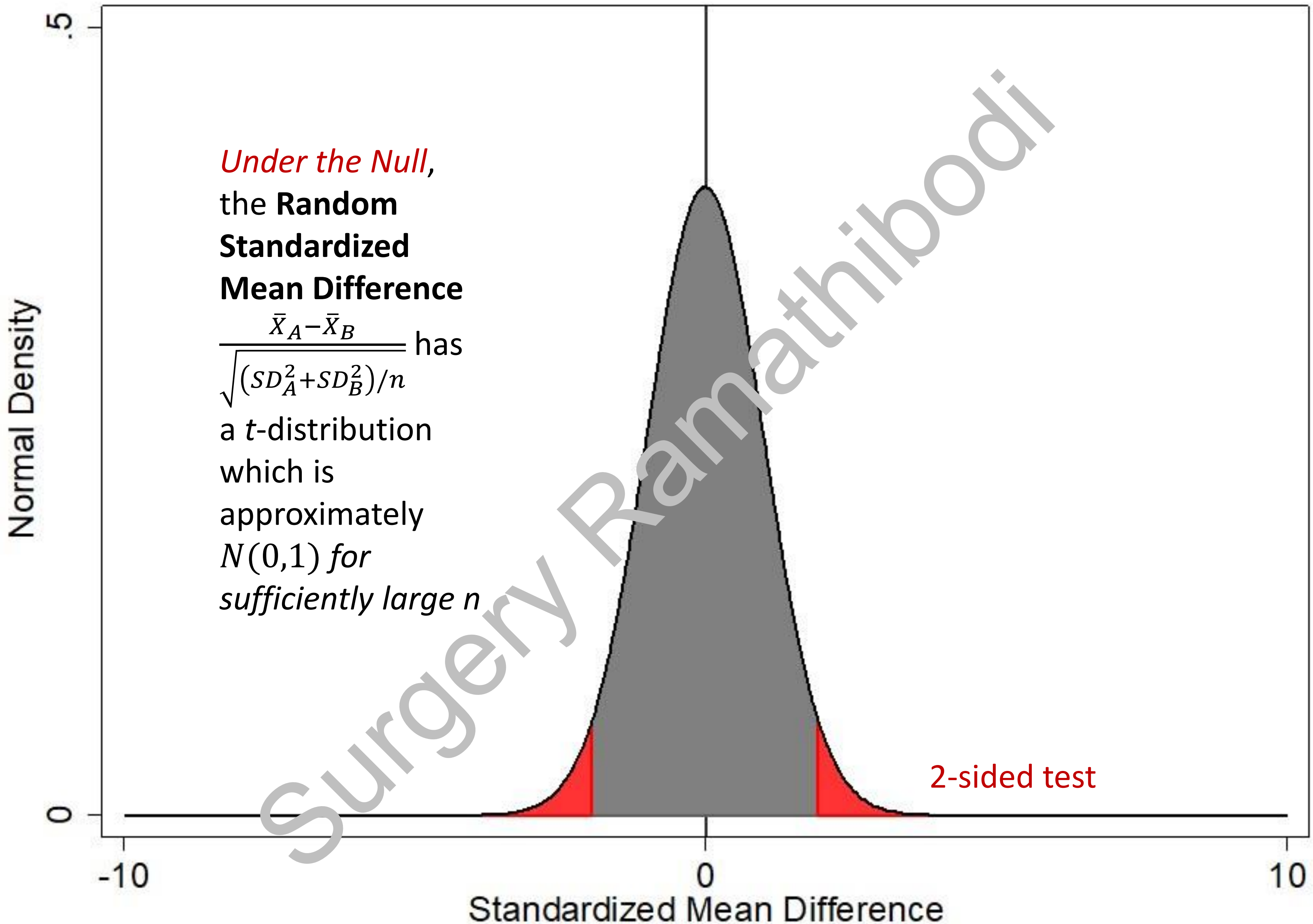
In Other Words …

- If a given true mean difference exists, say $\bar{x}_A - \bar{x}_B = 0.5$, then in *repeated experiments* the **statistical test will declare statistical significance** in at least 80 out of 100 trials *in the long run*
- This is, of course, purely *hypothetical*
- How can we use this idea to estimate a sample size?

The *Random* Mean Difference

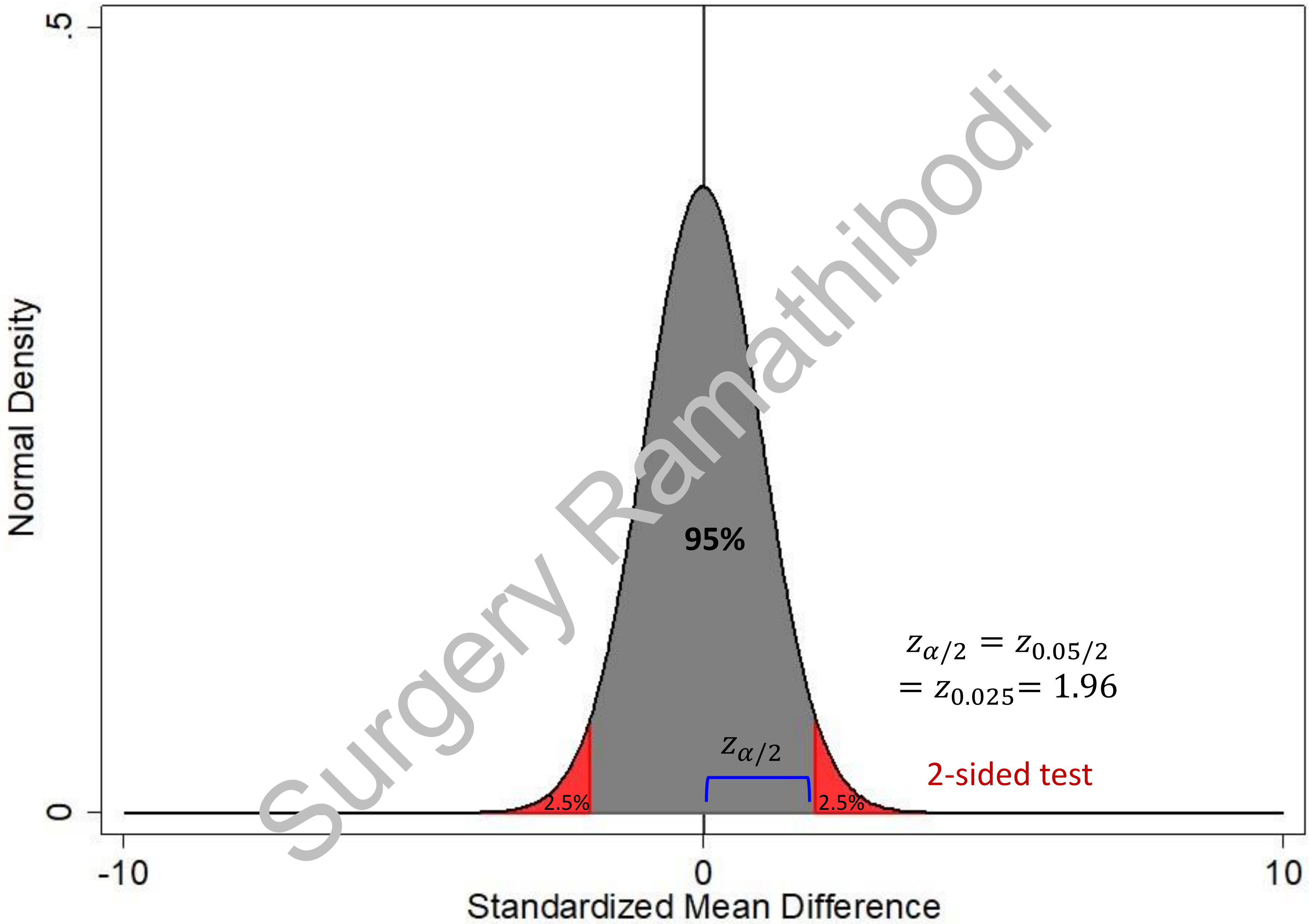
- The *Standardized Mean Difference*, under repeated experiments, will be a **random variable** with a probability distribution of values
- Under certain conditions, this probability distribution is approximately *Normal*
- Under the additional assumption of *no real difference* (**Null Hypothesis**) the distribution is *Standard Normal* $N(0,1)$

$$\frac{\bar{X}_A - \bar{X}_B}{\sqrt{\left(SD_A^2 + SD_B^2\right)/n}} \equiv T \approx Z$$

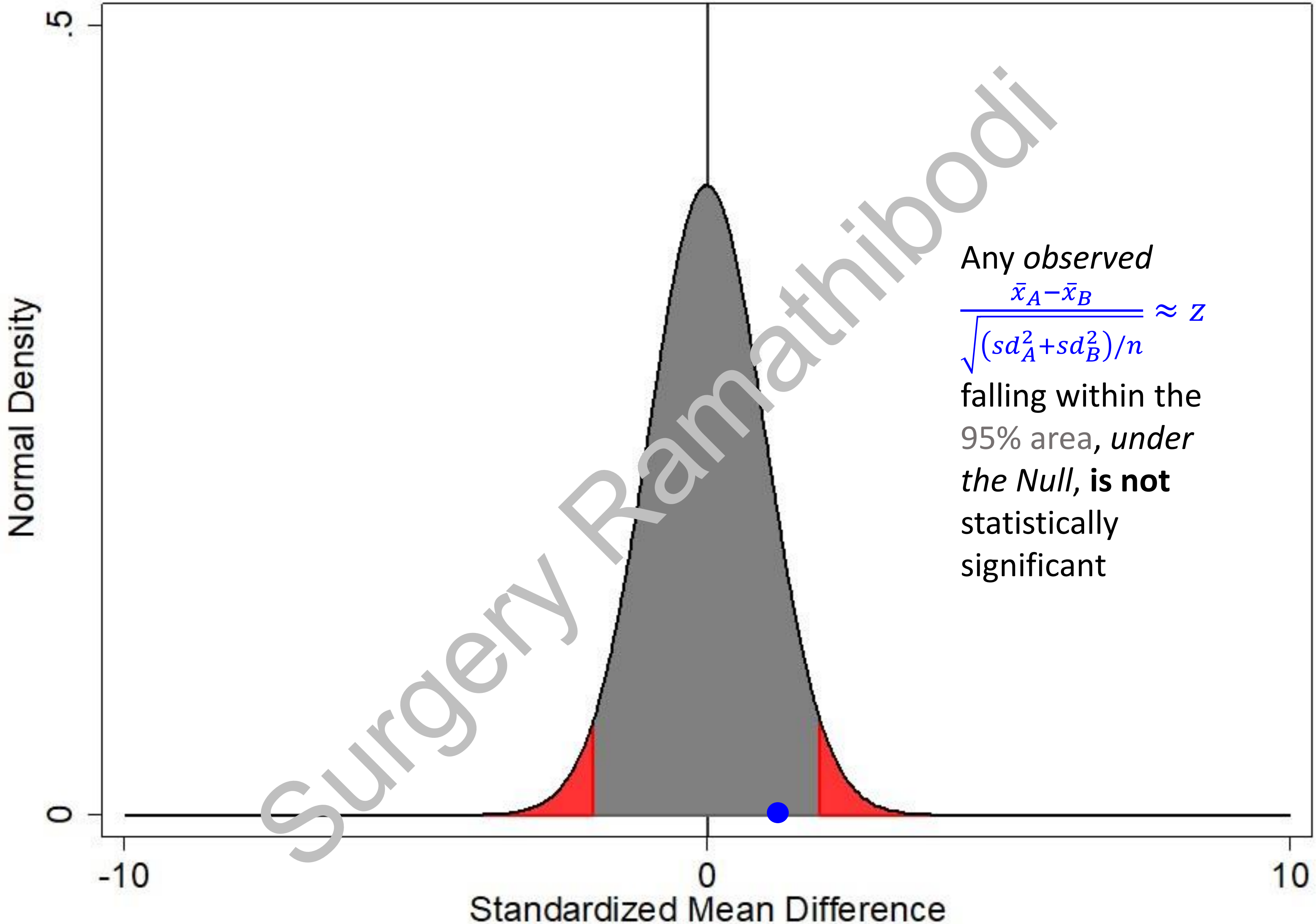


Under the Null,
the **Random Standardized Mean Difference**
$\frac{\bar{X}_A - \bar{X}_B}{\sqrt{(SD_A^2 + SD_B^2)/n}}$ has
a *t*-distribution which is approximately $N(0,1)$ *for sufficiently large n*

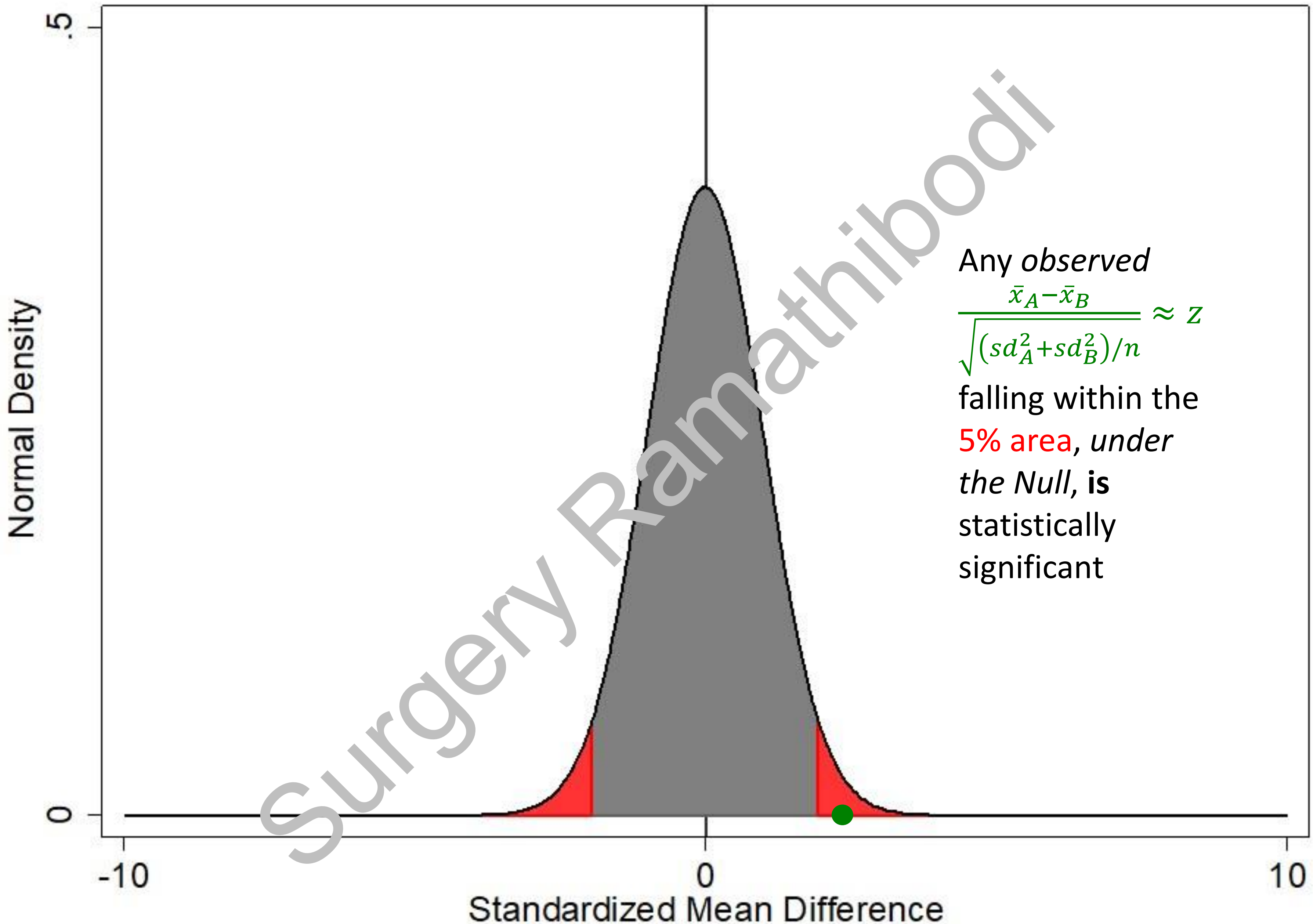

.5
0
Normal Density
95%
2.5%
2.5%
2-sided test
-10
0
10
Standardized Mean Difference



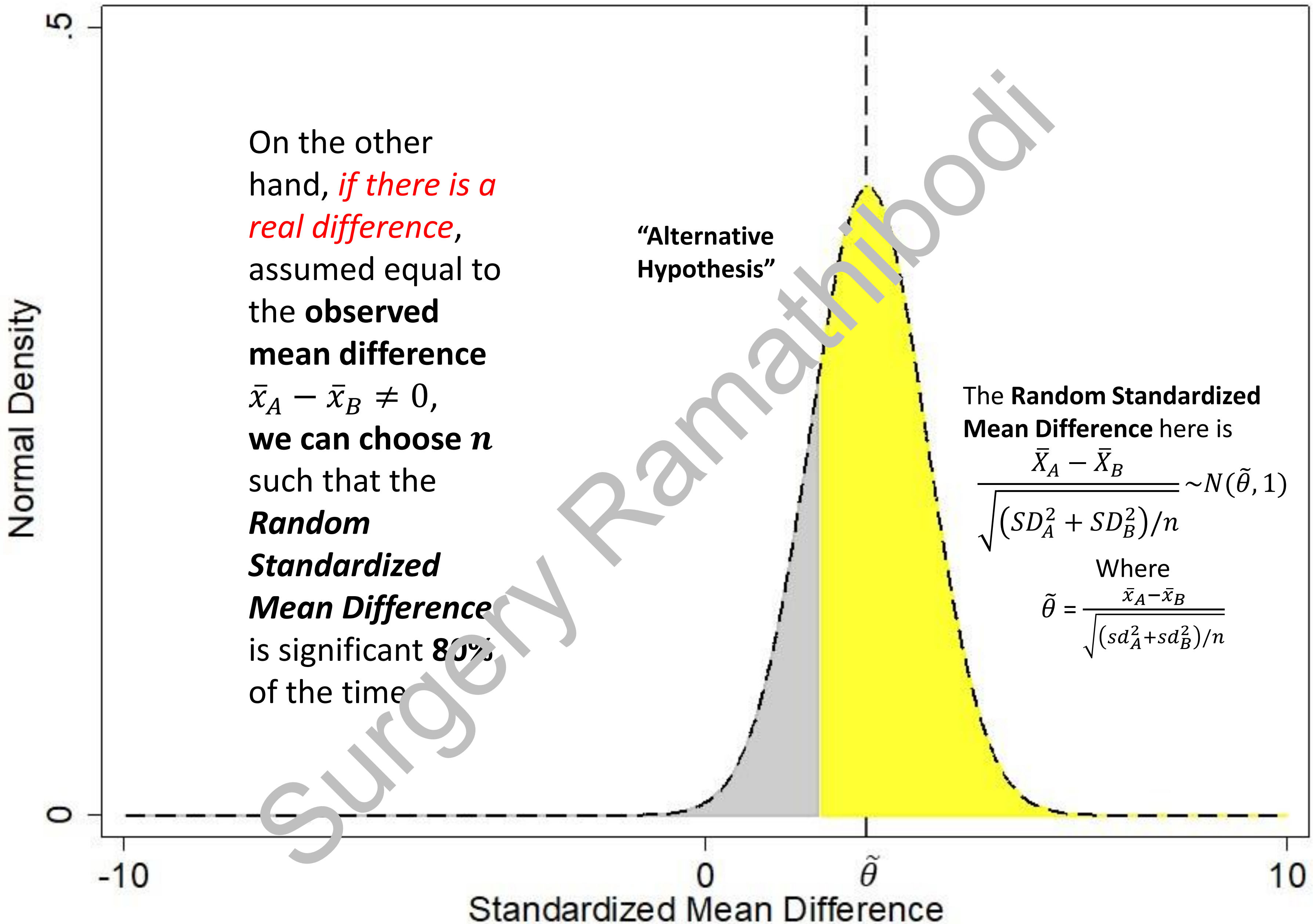
$$z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$$



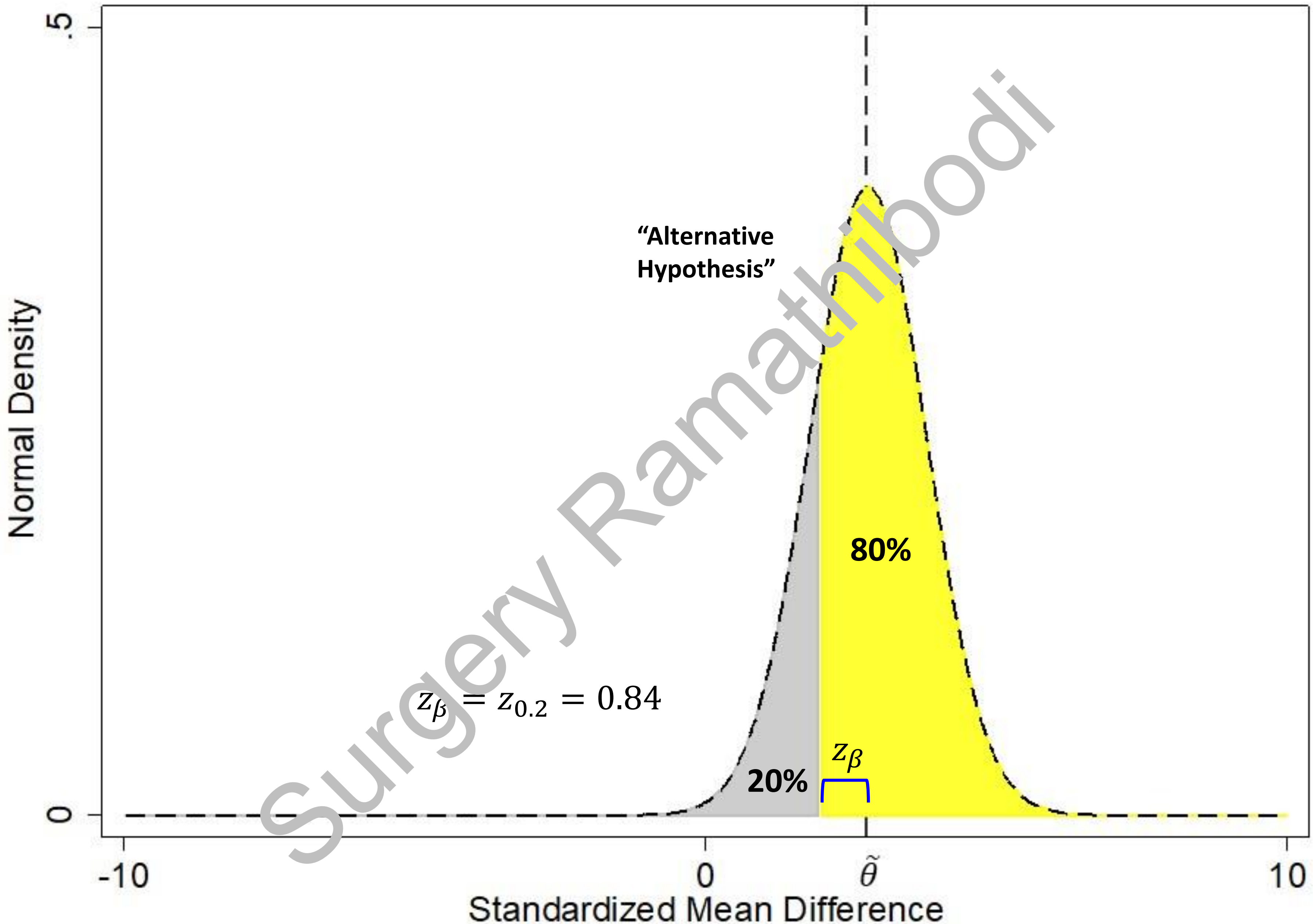
Any *observed* $\frac{\bar{x}_A - \bar{x}_B}{\sqrt{(sd_A^2 + sd_B^2)/n}} \approx z$ falling within the 95% area, *under the Null*, **is not** statistically significant

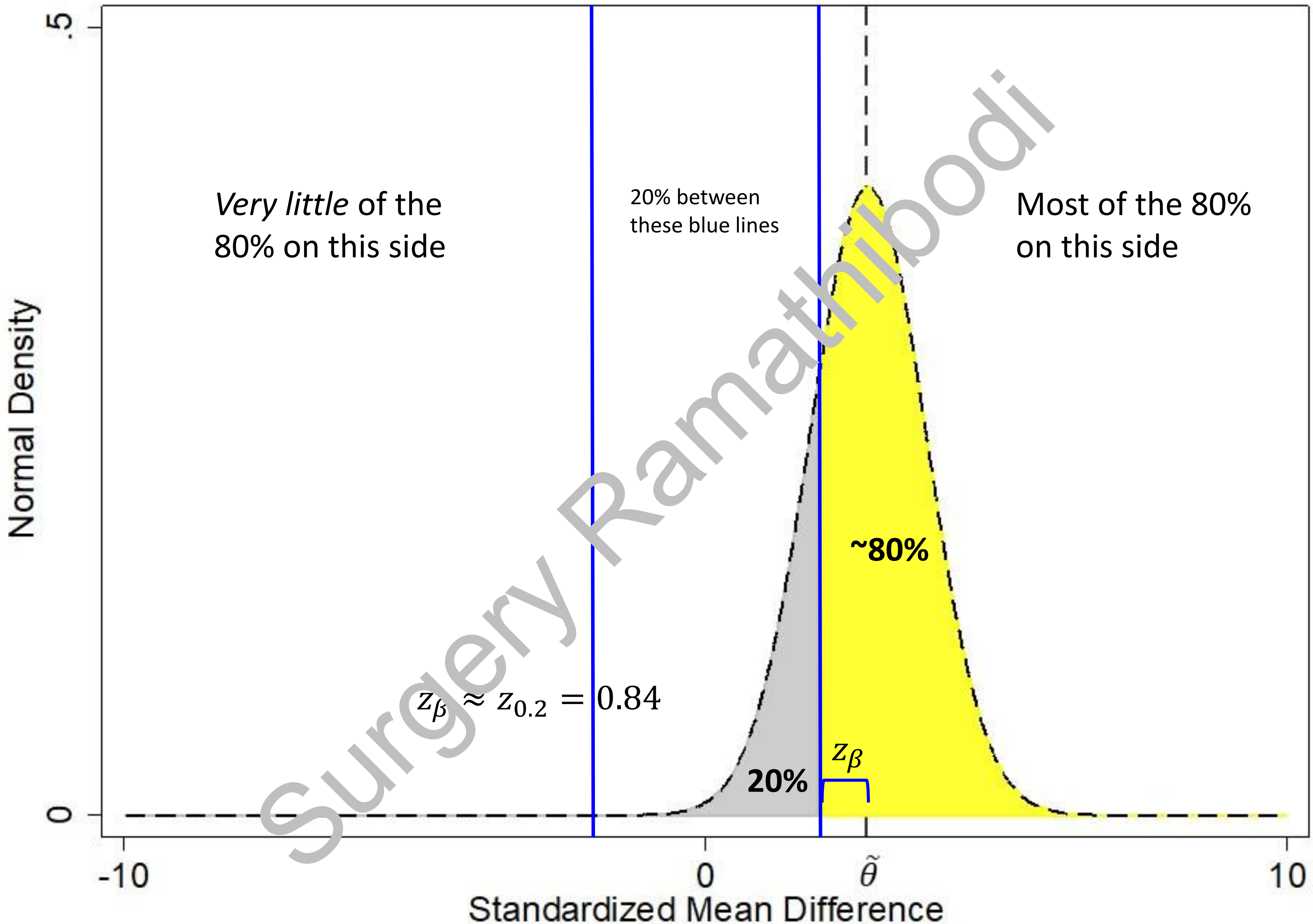


Any *observed* $\frac{\bar{x}_A - \bar{x}_B}{\sqrt{(sd_A^2 + sd_B^2)/n}} \approx z$ falling within the 5% area, *under the Null*, **is** statistically significant

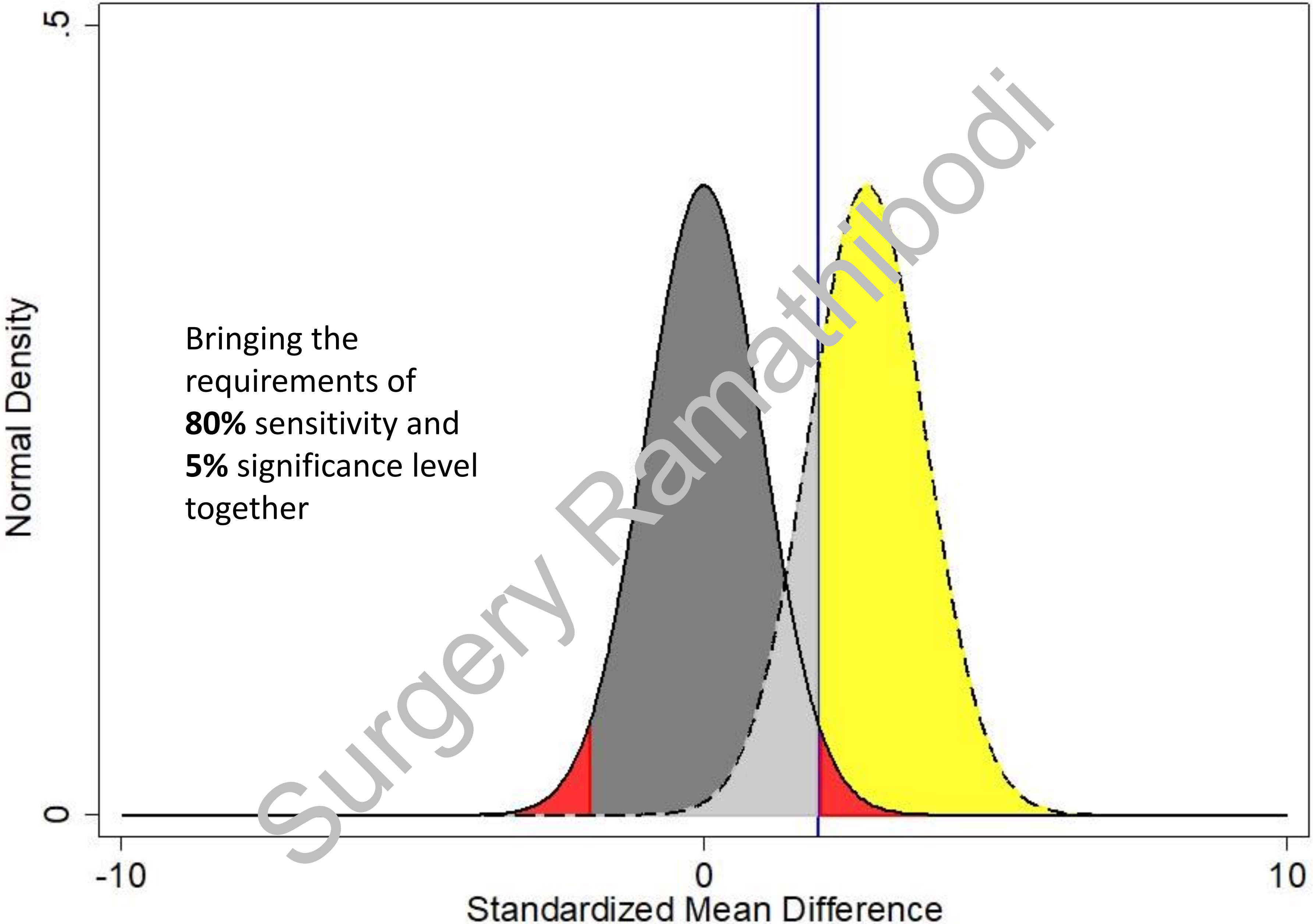


Illustrated here is the case where $\bar{x}_A - \bar{x}_B > 0$

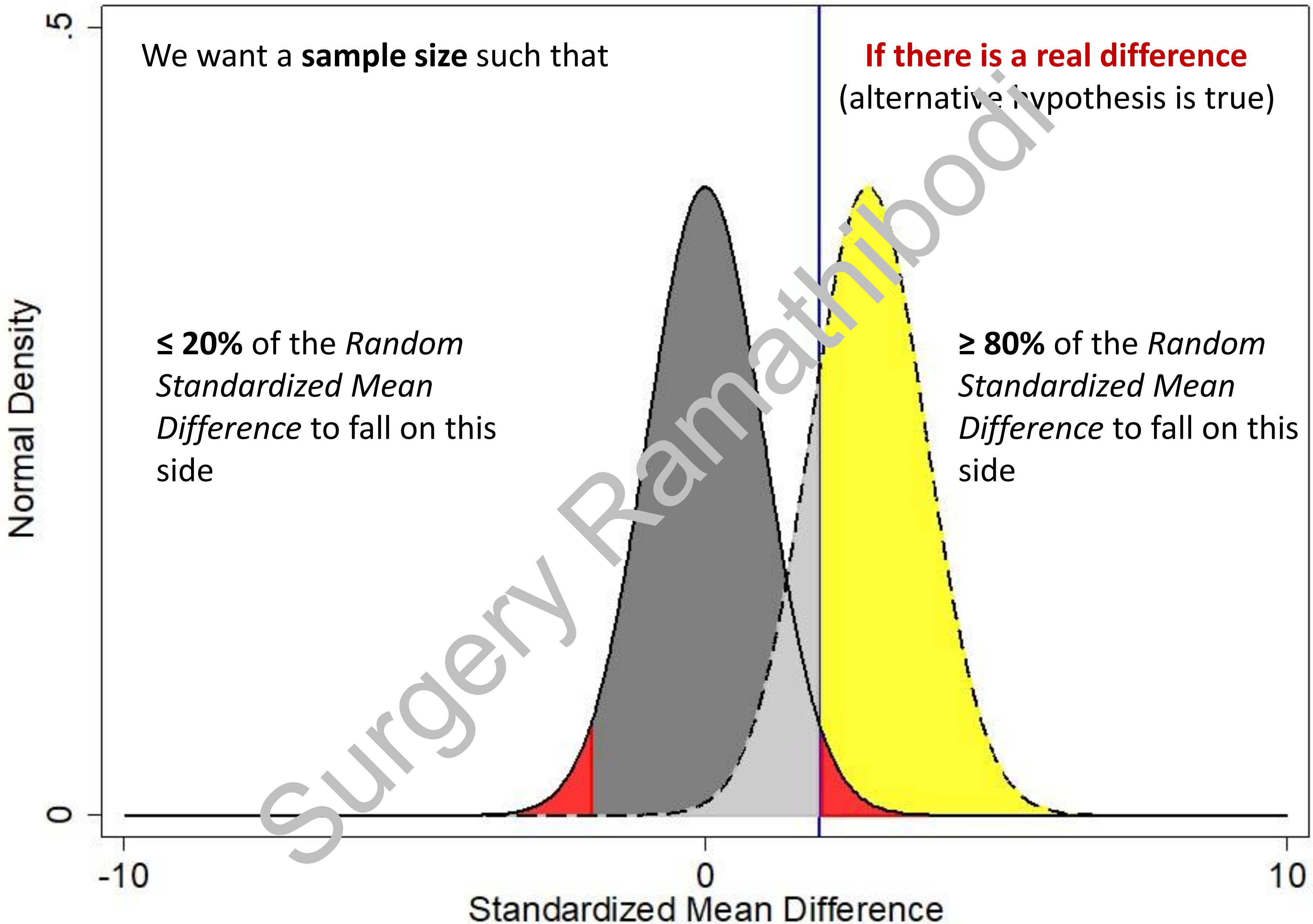

.5
0
Normal Density
"Alternative
Hypothesis"
80%
$z_\beta = z_{0.2} = 0.84$
z_β
20%
-10
0
$\tilde{\theta}$
10
Standardized Mean Difference

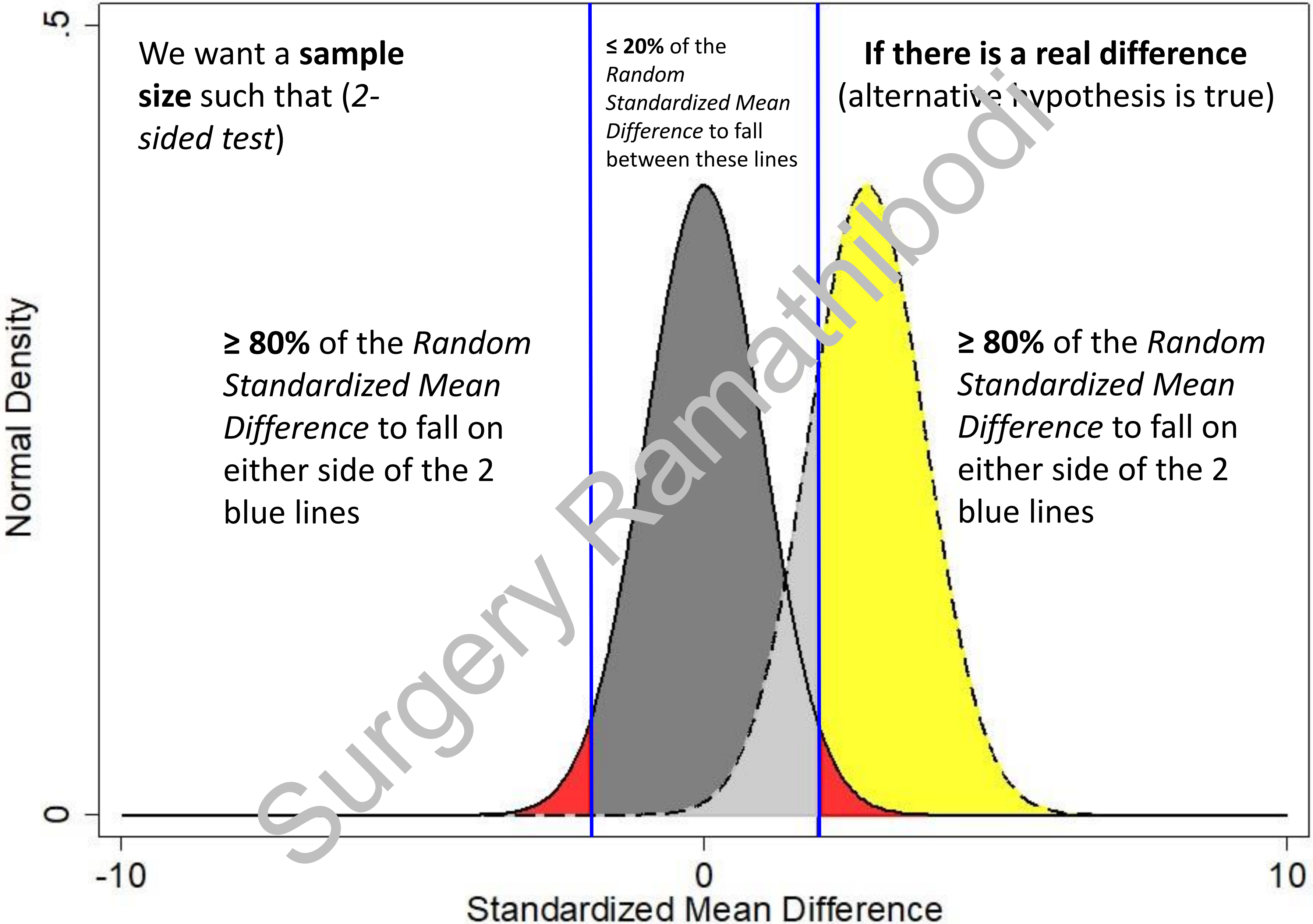


The actual situation for a 2-sided test

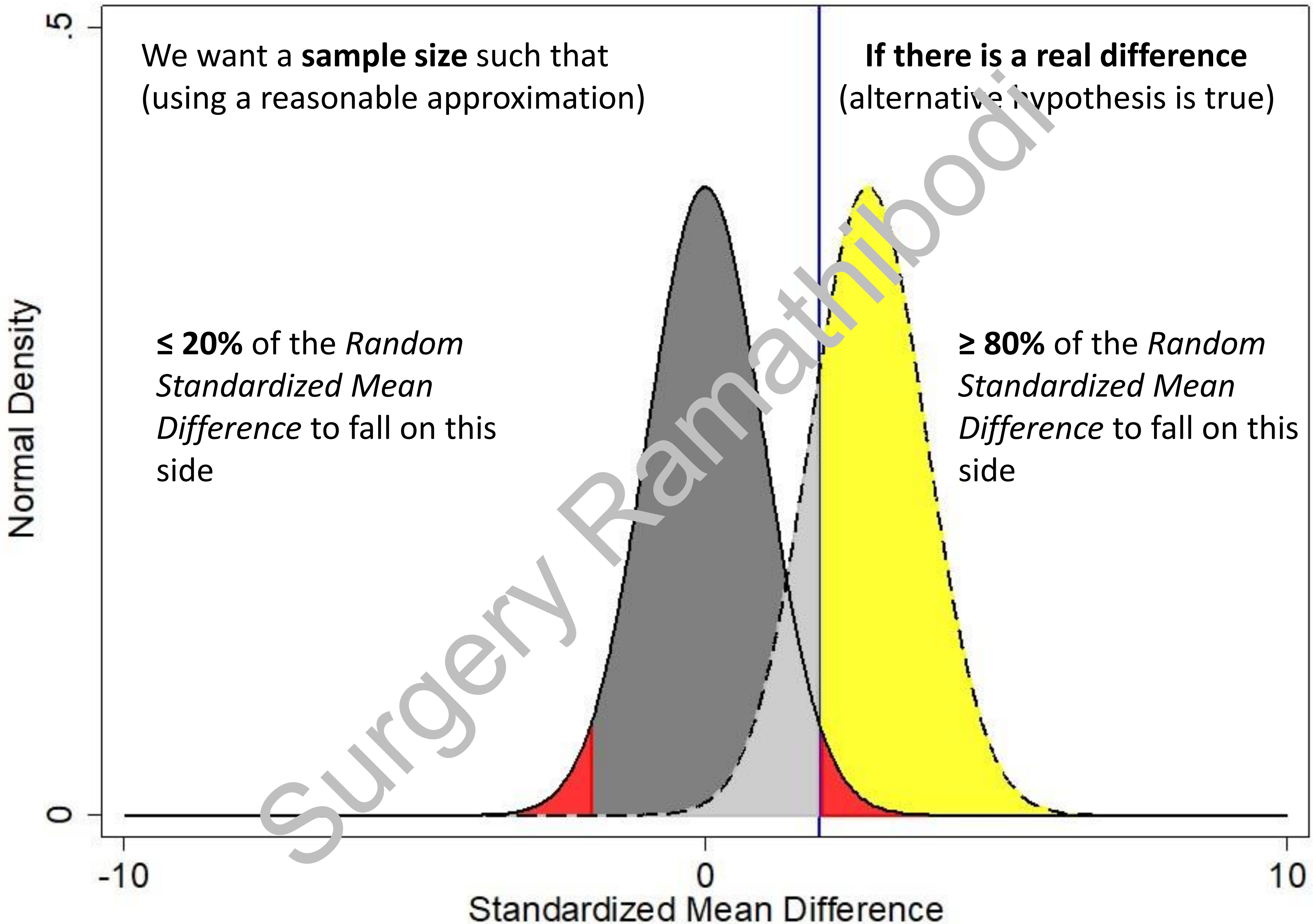


.5
0
Normal Density
-10
0
10
Standardized Mean Difference
Bringing the requirements of **80%** sensitivity and **5%** significance level together


We want a **sample size** such that
If there is a real difference
(alternative hypothesis is true)
≤ 20% of the *Random Standardized Mean Difference* to fall on this side
≥ 80% of the *Random Standardized Mean Difference* to fall on this side
Normal Density
.5
0
-10
0
10
Standardized Mean Difference

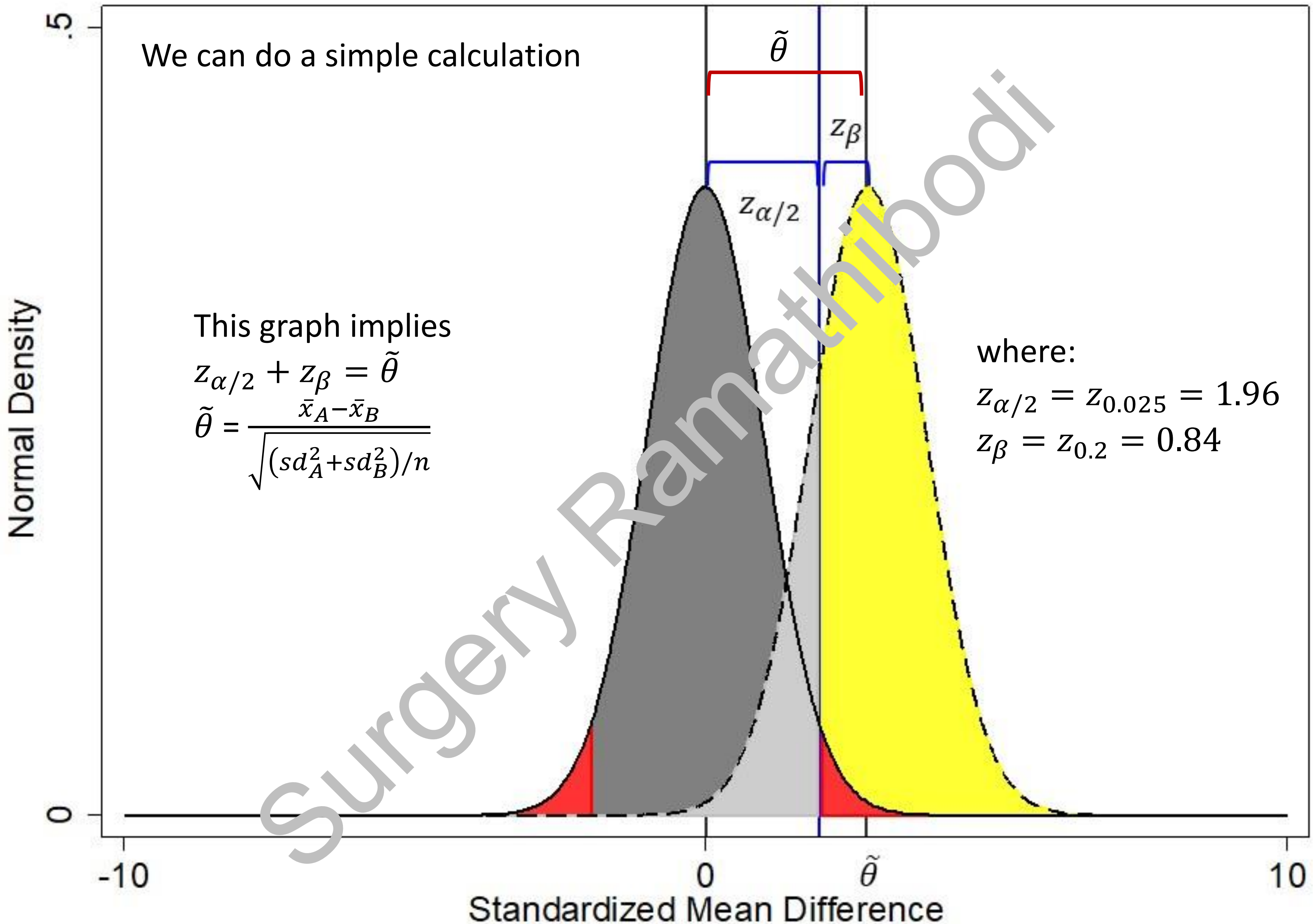


**This is the correct situation for a 2-sided test



**This is a reasonable approximate situation

We can do a simple calculation



This graph implies

$z_{\alpha/2} + z_{\beta} = \tilde{\theta}$

$\tilde{\theta} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(sd_A^2 + sd_B^2)/n}}$

where:

$z_{\alpha/2} = z_{0.025} = 1.96$

$z_{\beta} = z_{0.2} = 0.84$

The Sample Size Formula

Since $z_{\alpha/2} + z_{\beta} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(sd_A^2 + sd_B^2)/n}}$, therefore

$$\boxed{n = \frac{(z_{\alpha/2} + z_{\beta})^2 (sd_A^2 + sd_B^2)}{(\bar{x}_A - \bar{x}_B)^2} = \frac{(2.8)^2 (sd_A^2 + sd_B^2)}{(\bar{x}_A - \bar{x}_B)^2}}$$

And this is the prototypical sample size formula (for the "equality trial") – the *minimum* value

The Sample Size Formula

Since $$z_{\alpha/2} + z_{\beta} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(sd_A^2 + sd_B^2\right)/n}},$$ therefore

(1.96 + 0.84) = 2.8

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \left(sd_A^2 + sd_B^2\right)}{(\bar{x}_A - \bar{x}_B)^2} = \frac{(2.8)^2 \left(sd_A^2 + sd_B^2\right)}{(\bar{x}_A - \bar{x}_B)^2}$$

And this is the prototypical sample size formula (for the "equality trial") – the *minimum* value

Let’s Re-evaluate the Sample Size

- From the previous study
- We know the values of $\bar{x}_A, \bar{x}_B, sd_A, sd_B$
- We can re-evaluate the sample size required to detect this specific difference with 80% sensitivity (in statistics, sensitivity is called **Power**)
- The appropriate sample size should be, at least …
- $n = \frac{(2.8)^2(sd_A^2+sd_B^2)}{(\bar{x}_A-\bar{x}_B)^2} = \frac{(2.8)^2(1.5^2+1.6^2)}{(5.4-4.9)^2} = \cdots$

What if the Outcome is Binary?

- What if we are interested in pain level as the outcome?
- Suppose we wish to compare **severe pain** between the two groups, defined as VAS > 5
- From the previous study:

  Drug A: no. of patients with VAS > 5 = 4 (0.4)

  Drug B: no. of patients with VAS > 5 = 3 (0.3)
- Call these proportions p_A and p_B

What if the Outcome is Binary?

Replace

$$\bar{x}_A - \bar{x}_B \ with \ p_A - p_B$$

$$sd_A \ with \ \sqrt{p_A(1 - p_A)}$$

$$sd_B \ with \ \sqrt{p_B(1 - p_B)}$$

Sample Size Formula for Proportions

- The sample size formula becomes*

$$n = \frac{2.8^2\{p_A(1-p_A)+p_B(1-p_B)\}}{(p_A - p_B)^2}$$

- What is the estimated sample size for $p_A = 0.4$ and $p_B = 0.3$? (significance level 5%, power 80%)
- $n = \frac{2.8^2\{0.4\ (1-0.4)+0.3\ (1-0.3)\}}{(0.4-0.3)^2}$ =? What do you see?

*One of several reasonable approximate sample size formulae for proportions

More Exercises

- A clinical trial comparing preoperative antibiotics with no antibiotics in the prevention of postoperative infection
- No antibiotics: infection rate 20% (0.2)
- Calculate the sample size required …
- $n = \frac{2.8^2\{p_A(1-p_A)+p_B(1-p_B)\}}{(p_A-p_B)^2}$
- What is p_A? What is p_B?

Summary

We discussed/introduced

- Summary statistics
- Simple statistical tests
- Sample size estimation

Hopefully this is useful!

Thank you for Your Attention!