Variance estimation for the Kappa statistic in the presence of clustered data and heterogeneous observations

Mary M. Ryan

Dr. Daniel L. Gillen, Dr. William D. Spotnitz

University of California, Irvine,
for the ASQ Silicon Valley Section Statistics & Reliability Discussion Group

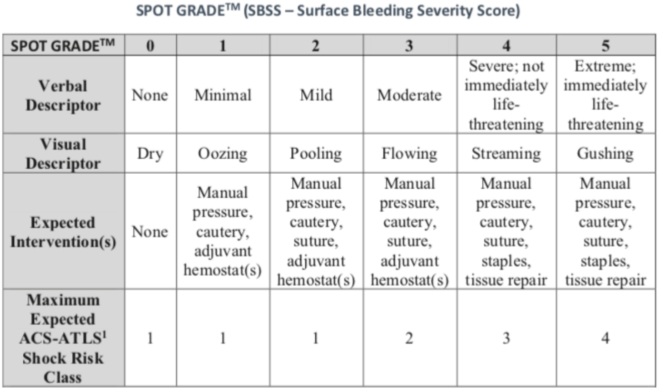
April 14, 2021

Motivation: SPOT GRADE Trial

- Researchers working on local hemostatic agent to stop bleeding on "low grade" wounds
- FDA required researchers to first develop scale to classify bleeds
  - Wanted surgeons to have better knowledge of what type of wounds appropriate to use agent on
  - Concerned surgeons would use agent on bleeds not be designed to stop

Motivation: SPOT GRADE Trial

- SPOT GRADE surface bleed severity scale (SBSS) developed to standardize severity of blood loss[13]
  - 6 categories: 0-5
  - Higher category ⇒ faster blood loss
  - Hemostatic agent designed for category 3 or lower



SPOT GRADE$^{\text{TM}}$ (SBSS – Surface Bleeding Severity Score)

SPOT GRADE$^{\text{TM}}$	0	1	2	3	4	5
Verbal Descriptor	None	Minimal	Mild	Moderate	Severe; not immediately life-threatening	Extreme; immediately life-threatening
Visual Descriptor	Dry	Oozing	Pooling	Flowing	Streaming	Gushing
Expected Intervention(s)	None	Manual pressure, cautery, adjuvant hemostat(s)	Manual pressure, cautery, suture, adjuvant hemostat(s)	Manual pressure, cautery, suture, adjuvant hemostat(s)	Manual pressure, cautery, suture, staples, tissue repair	Manual pressure, cautery, suture, staples, tissue repair
Maximum Expected ACS-ATLS[1] Shock Risk Class	1	1	1	2	3	4

Motivation: SPOT GRADE Trial

- Scores defined by flux/flow rate of blood from wound
- Higher flow rate ranges for larger scores and larger bleed surfaces
  - **Bleeds within same category can look very different**

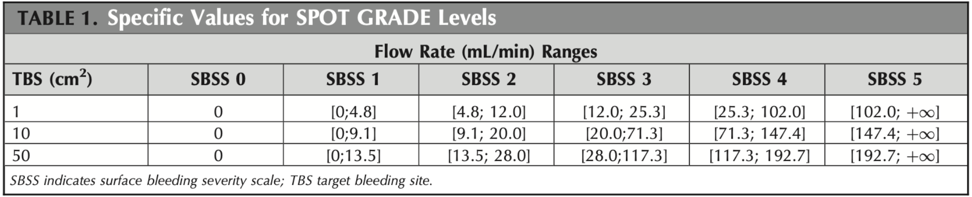


TABLE 1. Specific Values for SPOT GRADE Levels

	Flow Rate (mL/min) Ranges					
TBS (cm^2)	**SBSS 0**	**SBSS 1**	**SBSS 2**	**SBSS 3**	**SBSS 4**	**SBSS 5**
1	0	[0;4.8]	[4.8; 12.0]	[12.0; 25.3]	[25.3; 102.0]	[102.0; +∞]
10	0	[0;9.1]	[9.1; 20.0]	[20.0;71.3]	[71.3; 147.4]	[147.4; +∞]
50	0	[0;13.5]	[13.5; 28.0]	[28.0;117.3]	[117.3; 192.7]	[192.7; +∞]

SBSS indicates surface bleeding severity scale; TBS target bleeding site.

Motivation: SPOT GRADE Trial

- 14 surgeons watched video simulations in a randomized sequence and classified bleeding severity by SPOT GRADE category
  - 36 training videos
  - 36 testing videos
    - Each video viewed 3 times (108 total clips to view)

- **Kappa statistic** used to assess inter- and intra-rater reliability

Rating Data

- Rating data can be thought of as multinomial random variables:

$$(x_{11}, \ldots, x_{kk}) \sim \text{Multinomial}\left(N, [\pi_{11}, \ldots, \pi_{kk}]\right),$$

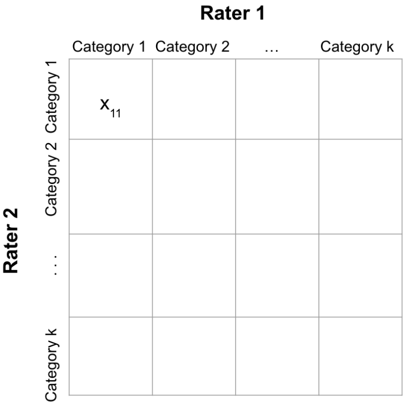
		Rater 1			
		Category 1	Category 2	…	Category k
Rater 2	Category 1	x_{11}			
	Category 2				
	…				
	Category k				

- How can we tell how well raters are agreeing with each other overall?

Rating Data

- Rating data can be thought of as multinomial random variables:

$$(x_{11}, \ldots, x_{kk}) \sim \text{Multinomial}\left(N, [\pi_{11}, \ldots, \pi_{kk}]\right),$$



Rater 1

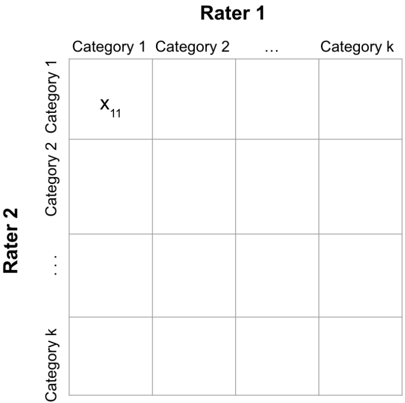
Rater 2	Category 1	Category 2	...	Category k
Category 1	x_{11}			
Category 2				
...				
Category k				

- How can we tell how well raters are agreeing with each other overall?
- Observed probability of agreement: $p_o = \sum_{i=1}^{k} p_{ii}$

Rating Data

- Rating data can be thought of as multinomial random variables:

$$(x_{11}, \ldots, x_{kk}) \sim \text{Multinomial}\left(N, [\pi_{11}, \ldots, \pi_{kk}]\right),$$



Rater 1

Rater 2

	Category 1	Category 2	...	Category k
Category 1	x_{11}			
Category 2				
...				
Category k				

- How can we tell how well raters are agreeing with each other overall?
- Observed probability of agreement: $p_o = \sum_{i=1}^{k} p_{ii}$
- Issue: expected probability of agreement by chance changes depending on marginal probability of classifying item to category
  - Can't just trust a "high" agreement probability to signal "high" agreement

Cohen's Kappa[2]

- Kappa statistic assesses **likelihood-above-chance** of two raters agreeing

$$\kappa = \frac{p_o - p_e}{1 - p_e} \in (-1, 1)$$

  - $p_o = \sum_{i=1}^{k} p_{ii}$
  - $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$
- $\kappa = 0$ implies rater agreement on par with chance
- $\kappa \rightarrow 1$ implies raters agree more
- $\kappa \rightarrow -1$ implies raters disagree more

Cohen's Kappa[2]

- Kappa statistic assesses **likelihood-above-chance** of two raters agreeing

$$\kappa = \frac{p_o - p_e}{1 - p_e} \in (-1, 1)$$

  - $p_o = \sum_{i=1}^{k} p_{ii}$
  - $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$
- $\kappa = 0$ implies rater agreement on par with chance
- $\kappa \rightarrow 1$ implies raters agree more
- $\kappa \rightarrow -1$ implies raters disagree more

- Assumes all items within a category are **exchangeable** and all ratings **independent**

Cohen's Kappa[2]

- Kappa statistic assesses **likelihood-above-chance** of two raters agreeing

$$\kappa = \frac{p_o - p_e}{1 - p_e} \in (-1, 1)$$

  - $p_o = \sum_{i=1}^{k} p_{ii}$
  - $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$
- $\kappa = 0$ implies rater agreement on par with chance
- $\kappa \rightarrow 1$ implies raters agree more
- $\kappa \rightarrow -1$ implies raters disagree more

- Assumes all items within a category are **exchangeable** and all ratings **independent**
- Landis & Koch's[9] interpretation of κ:

κ*value*	Interpretation
(-1, 0)	Poor agreement
[0, 0.2]	Slight agreement
(0.2, 0.4]	Fair agreement
(0.4, 0.6]	Moderate agreement
(0.6, 0.8]	Substantial agreement
(0.8, 1)	Almost perfect agreement

Cohen's Kappa[2]

- Kappa statistic assesses likelihood-above-chance of two raters agreeing

$$\kappa = \frac{p_o - p_e}{1 - p_e} \in (-1, 1)$$

  - $p_o = \sum_{i=1}^{k} p_{ii}$
  - $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$
- $\kappa = 0$ implies rater agreement on par with chance
- $\kappa \to 1$ implies raters agree more
- $\kappa \to -1$ implies raters disagree more
- Assumes all items within a category are **exchangeable** and all ratings **independent**
- Landis & Koch's[9] interpretation of κ:

κ*value*	Interpretation
(-1, 0)	Poor agreement
[0, 0.2]	Slight agreement
(0.2, 0.4]	Fair agreement
(0.4, 0.6]	Moderate agreement
(0.6, 0.8]	**Substantial agreement**
(0.8, 1)	Almost perfect agreement

Kappa Variations

- Weighted Kappa[3]: some misclassifications are a greater sin than others
  - $p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}$
  - $p_e = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.j}$
  - Quadratic weights: $w_{ij} = 1 - \frac{(i-j)^2}{(K-1)^2}$
  - Absolute weights: $w_{ij} = 1 - \frac{|i-j|}{(K-1)}$
- Kappa for multiple raters[4]
- Kappa for clustered data[8; 14; 16; 15]
- Using GEEs to incorporate rater and item covariate information into Kappa[6]
- And many more!

Kappa Asymptotics

- Fleiss et al.[5] asserted that, by CLT:

$$\sqrt{n}(\kappa - \kappa_0) \dot{\sim} N\left(0, \sigma^2_\kappa\right),$$

  where κ_0 is the true κ value, and σ^2_κ is a function of p_e, p_o, and n
- This means we can create confidence intervals and perform inference on κ

Kappa Asymptotics

- Fleiss et al.[5] asserted that, by CLT:

$$\sqrt{n}(\kappa - \kappa_0) \dot{\sim} N\left(0, \sigma^2_\kappa\right),$$

  where κ_0 is the true κ value, and σ^2_κ is a function of p_e, p_o, and n
- This means we can create confidence intervals and perform inference on κ
- Since $\kappa \in (-1, 1)$, Normal approximation from Fleiss et al. likely to perform poorly in small samples
- Propose transformation of κ to map onto $\mathbb{R}$:

$$f(\kappa) = ln\left(\frac{1+\kappa}{1-\kappa}\right) \equiv \varphi$$

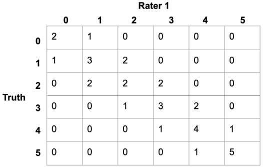
  - Can calculate CI for φ then back-transform to put it on regular κ scale for interpretation

SPOT GRADE Data

How are we using Kappa in the SPOT GRADE study?

SPOT GRADE Data

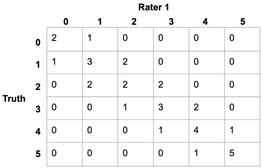
How are we using Kappa in the SPOT GRADE study?



		Rater 1					
		0	1	2	3	4	5
Truth	0	2	1	0	0	0	0
	1	1	3	2	0	0	0
	2	0	2	2	2	0	0
	3	0	0	1	3	2	0
	4	0	0	0	1	4	1
	5	0	0	0	0	1	5

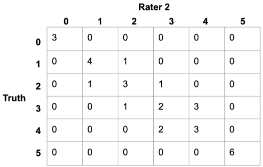
SPOT GRADE Data

How are we using Kappa in the SPOT GRADE study?



		Rater 1					
		0	1	2	3	4	5
Truth	0	2	1	0	0	0	0
	1	1	3	2	0	0	0
	2	0	2	2	2	0	0
	3	0	0	1	3	2	0
	4	0	0	0	1	4	1
	5	0	0	0	0	1	5

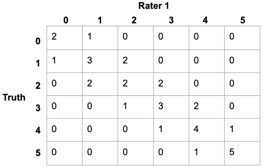
+



		Rater 2					
		0	1	2	3	4	5
Truth	0	3	0	0	0	0	0
	1	0	4	1	0	0	0
	2	0	1	3	1	0	0
	3	0	0	1	2	3	0
	4	0	0	0	2	3	0
	5	0	0	0	0	0	6

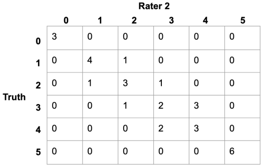
SPOT GRADE Data

How are we using Kappa in the SPOT GRADE study?



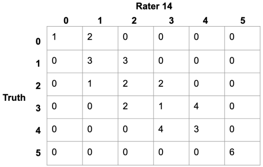
		Rater 1					
		0	1	2	3	4	5
Truth	0	2	1	0	0	0	0
	1	1	3	2	0	0	0
	2	0	2	2	2	0	0
	3	0	0	1	3	2	0
	4	0	0	0	1	4	1
	5	0	0	0	0	1	5

$+$



		Rater 2					
		0	1	2	3	4	5
Truth	0	3	0	0	0	0	0
	1	0	4	1	0	0	0
	2	0	1	3	1	0	0
	3	0	0	1	2	3	0
	4	0	0	0	2	3	0
	5	0	0	0	0	0	6

$+ \dots +$



		Rater 14					
		0	1	2	3	4	5
Truth	0	1	2	0	0	0	0
	1	0	3	3	0	0	0
	2	0	1	2	2	0	0
	3	0	0	2	1	4	0
	4	0	0	0	4	3	0
	5	0	0	0	0	0	6

SPOT GRADE Data

How are we using Kappa in the SPOT GRADE study?

Rater 1

Truth	0	1	2	3	4	5
0	2	1	0	0	0	0
1	1	3	2	0	0	0
2	0	2	2	2	0	0
3	0	0	1	3	2	0
4	0	0	0	1	4	1
5	0	0	0	0	1	5

+

Rater 2

Truth	0	1	2	3	4	5
0	3	0	0	0	0	0
1	0	4	1	0	0	0
2	0	1	3	1	0	0
3	0	0	1	2	3	0
4	0	0	0	2	3	0
5	0	0	0	0	0	6

+ … +

Rater 14

Truth	0	1	2	3	4	5
0	1	2	0	0	0	0
1	0	3	3	0	0	0
2	0	1	2	2	0	0
3	0	0	2	1	4	0
4	0	0	0	4	3	0
5	0	0	0	0	0	6

=

Raters

Truth	0	1	2	3	4	5
0	40	2	0	0	0	0
1	2	76	8	0	0	0
2	0	8	70	14	0	0
3	0	0	14	72	12	0
4	0	0	0	12	78	1
5	0	0	0	0	1	81

Performing Kappa on the additive rating table to assess how reliable surgeons are at correctly classifying videos

Issues & Goals

- Rating same video multiple times induces **clustering** that biases variance estimate

Issues & Goals

- Rating same video multiple times induces **clustering** that biases variance estimate
- Videos within same category might not all have same probability of correct classification
  - Different combinations of surface area and flow rate
  - Operating characteristics of Kappa's asymptotic variance not yet explored under this setting

Issues & Goals

- Rating same video multiple times induces **clustering** that biases variance estimate
- Videos within same category might not all have same probability of correct classification
    - Different combinations of surface area and flow rate
    - Operating characteristics of Kappa's asymptotic variance not yet explored under this setting
- **Goal: Want to adapt Kappa statistic for clustered data and heterogeneity within categories by correcting variance estimate**

Simulated Data Generation

- 2 kinds of item heterogeneity we're dealing with here that we need simulated data to reflect:
  - Some SBSS categories are inherently easier (0, 5) or more difficult (2, 3) to correctly place than others (**between-category heterogeneity**)
  - Some videos within an SBSS category may be easier/more difficult to correctly place than others (**within-category heterogeneity**)
- How do we incorporate these into video classification probabilities?

Simulated Data Generation

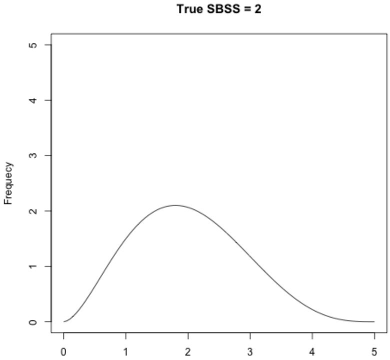
- Let π_{hmj} be the probability video j classified as category m when actually category h

Simulated Data Generation

- Let π_{hmj} be the probability video j classified as category m when actually category h
- $\pi_{hmj} = \int_{m-0.5}^{m+0.5} \frac{\left(\frac{1}{5}u\right)^{\alpha_{hj}-1}(1-\frac{1}{5}u)^{\beta_{hj}-1}\Gamma(\alpha_{hj}+\beta_{hj})}{5\Gamma(\alpha_{hj})\Gamma(\beta_{hj})} du$
- $\frac{\alpha_{hj}}{\alpha_{hj}+\beta_{hj}} \times 5 = h$
- $log(\beta_{hj}) \overset{indep.}{\sim} N(\mu_h, \sigma_h^2)$
- $\alpha_{hj} = \frac{\beta_{hj}h}{5-h}$

Simulated Data Generation

- Let π_{hmj} be the probability video j classified as category m when actually category h
- $\pi_{hmj} = \int_{m-0.5}^{m+0.5} \frac{\left(\frac{1}{5}u\right)^{\alpha_{hj}-1}(1-\frac{1}{5}u)^{\beta_{hj}-1}\Gamma(\alpha_{hj}+\beta_{hj})}{5\Gamma(\alpha_{hj})\Gamma(\beta_{hj})} du$
- $\frac{\alpha_{hj}}{\alpha_{hj}+\beta_{hj}} \times 5 = h$
- $log(\beta_{hj}) \overset{indep.}{\sim} N(\mu_h, \sigma_h^2)$
- $\alpha_{hj} = \frac{\beta_{hj} h}{5-h}$
- μ_h controls probability of correct classification
- σ_h^2 is increased or decreased to create random video effects for each unique video



$\mu_2 = 2.7, \ \sigma_2^2 = 1$

Variance Bias: Unclustered Data

Let's see how Kappa behaves under video heterogeneity, but no clustering

- 10,000 simulations
- N=14 surgeons per simulation
- Three Kappa values: 0.4, 0.6, 0.8
- Four video heterogeneity settings:

Heterogeneity Level	SBSS Category					
	0	1	2	3	4	5
None	0	0	0	0	0	0
Low	0.25	0.5	1	1	0.5	0.25
Medium	0.5	1	2	2	1	0.5
High	1	2	3	3	2	1

- **18 videos** per SBSS category, each **rated once** per surgeon

Variance Bias: Unclustered Data

- $\text{Variance ratio} = \frac{\text{Analytic variance}}{\text{Empirical variance}}$

		Video Heterogeneity
		None
$\kappa = 0.4$	Variance Ratio	1.127
	Coverage	0.963
$\kappa = 0.6$	Variance Ratio	1.125
	Coverage	0.960
$\kappa = 0.8$	Variance Ratio	1.061
	Coverage	0.952

Variance Bias: Unclustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

		Video Heterogeneity	
		None	Low
$\kappa = 0.4$	Variance Ratio	1.127	1.143
	Coverage	0.963	0.963
$\kappa = 0.6$	Variance Ratio	1.125	1.202
	Coverage	0.960	0.969
$\kappa = 0.8$	Variance Ratio	1.061	1.221
	Coverage	0.952	0.970

Variance Bias: Unclustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

		Video Heterogeneity			
		None	Low	Medium	High
$\kappa = 0.4$	Variance Ratio	1.127	1.143	1.306	1.672
	Coverage	0.963	0.963	0.974	0.989
$\kappa = 0.6$	Variance Ratio	1.125	1.202	1.392	1.736
	Coverage	0.960	0.969	0.979	0.991
$\kappa = 0.8$	Variance Ratio	1.061	1.221	1.682	2.181
	Coverage	0.952	0.970	0.988	0.997

- **Analytic variance is inflated**
- Increasing within-category video heterogeneity exacerbates this

Variance Bias: Clustered Data

Does adding clustering change the previous results we saw?

- 10,000 simulations
- n=50 surgeons per simulation
- Three Kappa values: 0.4, 0.6, 0.8
- Four video heterogeneity settings:

Heterogeneity Level	SBSS Category					
	0	1	2	3	4	5
None	0	0	0	0	0	0
Low	0.25	0.5	1	1	0.5	0.25
Medium	0.5	1	2	2	1	0.5
High	1	2	3	3	2	1

- **Six videos** per SBSS category, each **rated three times** per surgeon

Variance Bias: Clustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

		Video Heterogeneity
		None
$\kappa = 0.4$	Est. Kappa	0.404
	Variance Ratio	1.130
	Coverage	0.961
$\kappa = 0.6$	Est. Kappa	0.604
	Variance Ratio	1.146
	Coverage	0.961
$\kappa = 0.8$	Est. Kappa	0.795
	Variance Ratio	1.067
	Coverage	0.958

Variance Bias: Clustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

		Video Heterogeneity	
		None	Low
$\kappa = 0.4$	Est. Kappa	0.404	0.474
	Variance Ratio	1.130	1.186
	Coverage	0.961	0.963
$\kappa = 0.6$	Est. Kappa	0.604	0.585
	Variance Ratio	1.146	1.253
	Coverage	0.961	0.971
$\kappa = 0.8$	Est. Kappa	0.795	0.747
	Variance Ratio	1.067	1.228
	Coverage	0.958	0.968

- Increases of video heterogeneity, combined with data clustering, inflates analytic variance - not much different than we saw without clustering

Variance Bias: Clustered Data

- Variance ratio $= \frac{\text{Analytic variance}}{\text{Empirical variance}}$

		Video Heterogeneity			
		None	Low	Medium	High
$\kappa = 0.4$	Est. Kappa	0.404	0.474	0.419	0.477
	Variance Ratio	1.130	1.186	1.354	1.555
	Coverage	0.961	0.963	0.977	0.985
$\kappa = 0.6$	Est. Kappa	0.604	0.585	0.616	0.606
	Variance Ratio	1.146	1.253	1.329	1.900
	Coverage	0.961	0.971	0.978	0.993
$\kappa = 0.8$	Est. Kappa	0.795	0.747	0.795	0.825
	Variance Ratio	1.067	1.228	1.494	2.036
	Coverage	0.958	0.968	0.984	0.995

- Increases of video heterogeneity, combined with data clustering, inflates analytic variance - not much different than we saw without clustering
- May **bootstrap** new variance estimate to correct this

Variance Bias: Bootstrap

- Sampling units are surgeons, not videos
- Each bootstrap iteration will sample *n* surgeons

Algorithm 1: Bootstrap algorithm for variance of Kappa statistic.

```
for $b$ in $B$ do
    Randomly choose $n$ surgeons, with replacement;
    Take all observations belonging to sampled surgeons, and place in one
      contingency table;
    Find statistic, $\kappa_b$;
    Transform $\kappa_b$ to $\varphi_b$;
end
Calculate $\bar{\varphi} = \frac{1}{B}\sum_{b=1}^{B} \varphi_b$;
Calculate $\hat{\sigma}_B^2 = \frac{\sum_{b=1}^{B}(\varphi_b - \bar{\varphi})^2}{B-1}$
```

Variance Bias: Bootstrap

- Sampling units are surgeons, not videos
- Each bootstrap iteration will sample n surgeons

Algorithm 1: Bootstrap algorithm for variance of Kappa statistic.

```
for b in B do
    Randomly choose $n$ surgeons, with replacement;
    Take all observations belonging to sampled surgeons, and place in one
      contingency table;
    Find statistic, $\kappa_b$;
    Transform $\kappa_b$ to $\varphi_b$;
end
Calculate $\bar{\varphi} = \frac{1}{B}\sum_{b=1}^{B}\varphi_b$;
Calculate $\hat{\sigma}_B^2 = \frac{\sum_{b=1}^{B}(\varphi_b - \bar{\varphi})^2}{B-1}$
```

- Use $\hat{\sigma}_B^2$ instead of analytic variance estimate

Variance Bias: Clustered Data

- Employing bootstrap (200 samples) attenuates variance ratio back toward 1:

		Video Heterogeneity
		None
$\kappa = 0.4$	Est. Kappa	0.404
	Variance Ratio	0.984
	Coverage	0.940
$\kappa = 0.6$	Est. Kappa	0.604
	Variance Ratio	0.993
	Coverage	0.947
$\kappa = 0.8$	Est. Kappa	0.795
	Variance Ratio	0.965
	Coverage	0.938

Variance Bias: Clustered Data

- Employing bootstrap (200 samples) attenuates variance ratio back toward 1:

		Video Heterogeneity	
		None	Low
$\kappa = 0.4$	Est. Kappa	0.404	0.413
	Variance Ratio	0.984	1.009
	Coverage	0.940	0.942
$\kappa = 0.6$	Est. Kappa	0.604	0.599
	Variance Ratio	0.993	0.983
	Coverage	0.947	0.942
$\kappa = 0.8$	Est. Kappa	0.795	0.758
	Variance Ratio	0.965	0.962
	Coverage	0.938	0.937

Variance Bias: Clustered Data

- Employing bootstrap (200 samples) attenuates variance ratio back toward 1:

		Video Heterogeneity			
		None	Low	Medium	High
$\kappa = 0.4$	Est. Kappa	0.404	0.413	0.438	0.578
	Variance Ratio	0.984	1.009	0.971	0.973
	Coverage	0.940	0.942	0.940	0.940
$\kappa = 0.6$	Est. Kappa	0.604	0.599	0.652	0.679
	Variance Ratio	0.993	0.983	1.004	0.979
	Coverage	0.947	0.942	0.942	0.937
$\kappa = 0.8$	Est. Kappa	0.795	0.758	0.726	0.721
	Variance Ratio	0.965	0.962	0.983	0.994
	Coverage	0.938	0.937	0.939	0.941

- Bootstrap procedure corrects variance overestimation
- Slight undercoverage happening

Application to SPOT GRADE: Presence of Heterogeneity

- Fixed variance bias in simulation

Application to SPOT GRADE: Presence of Heterogeneity

- Fixed variance bias in simulation

🥳

Application to SPOT GRADE: Presence of Heterogeneity

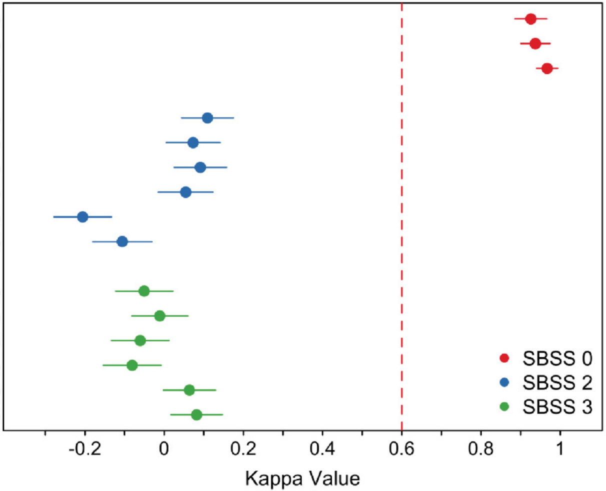
- Fixed variance bias in simulation

🥳

- Is heterogeneity actually a problem in real studies?
  - Do we see between-category heterogeneity? Within-category heterogeneity?
- Compared surgeons' ability to correctly classify **individual videos** within the same category vs. all videos in a reference category(s) using Kappa
  - If within-category kappas varied lots ⇒ lots of **within-category hetereogeneity**
  - If kappas between categories varied lots ⇒ lots of **between-category hetereogeneity**

Application to SPOT GRADE: Presence of Heterogeneity

	Video	Kappa	(95% CI)
0 vs Other	1	0.93	(0.89 , 0.97)
	2	0.94	(0.90 , 0.97)
	3	0.97	(0.94 , 0.99)
2 vs All 3	1	0.11	(0.04 , 0.17)
	2	0.07	(0.01 , 0.14)
	3	0.09	(0.02 , 0.16)
	4	0.05	(-0.01 , 0.12)
	5	-0.21	(-0.28 , -0.13)
	6	-0.11	(-0.18 , -0.03)
3 vs All 2	1	-0.05	(-0.12 , 0.02)
	2	-0.01	(-0.08 , 0.06)
	3	-0.06	(-0.13 , 0.01)
	4	-0.08	(-0.15 , -0.01)
	5	0.06	(0.00 , 0.13)
	6	0.08	(0.02 , 0.15)



Application to SPOT GRADE: Identification of Eligibility

- For development later clinical trial of local hemostatic device, important to be able to identify study-eligible bleeds (SBSS 1-3) from study-ineligible bleeds (SBSS 4-5)
- Testing hypothesis

$$H_0: \ \kappa_E \leq 0.60 \quad \text{vs.} \quad H_1: \kappa_E > 0.60$$

Application to SPOT GRADE: Identification of Eligibility

- For development later clinical trial of local hemostatic device, important to be able to identify study-eligible bleeds (SBSS 1-3) from study-ineligible bleeds (SBSS 4-5)
- Testing hypothesis

$$H_0 : \ \kappa_E \leq 0.60 \quad \text{vs.} \quad H_1 : \kappa_E > 0.60$$

Partial Z Transformation	Bootstrapped Variance Est.	κ (95% CI)
✓	✓	0.811 (0.810, 0.813)
✓	✗	0.811 (0.791, 0.830)
✗	✗	0.833 (0.806, 0.861)

Application to SPOT GRADE: Identification of Hemostasis

- To accurately assess whether the local hemostatic device under consideration was effective, necessary for surgeons to be able to identify whether hemostasis had been achieved (SBSS 0) or not (SBSS > 0).
- Testing hypothesis

$$H_0 : \ \kappa_H \leq 0.60 \quad \text{vs.} \quad H_1 : \kappa_H > 0.60$$

Application to SPOT GRADE: Identification of Hemostasis

- To accurately assess whether the local hemostatic device under consideration was effective, necessary for surgeons to be able to identify whether hemostasis had been achieved (SBSS 0) or not (SBSS > 0).
- Testing hypothesis

$$H_0: \ \kappa_H \leq 0.60 \quad \text{vs.} \quad H_1: \kappa_H > 0.60$$

Partial Z Transformation	Bootstrapped Variance Est.	κ (95% CI)
✓	✓	0.954 (0.952, 0.955)
✓	✗	0.954 (0.947, 0.960)
✗	✗	0.952 (0.930, 0.973)

Conclusions

- Even with slight amounts of variability among classification probabilities within categories, Kappa's analytic variance largely overestimates the true variance
    - Application of the bootstrap corrects for this overestimation, allowing for the correct inference of the Kappa statistic

Conclusions

- Even with slight amounts of variability among classification probabilities within categories, Kappa's analytic variance largely overestimates the true variance
  - Application of the bootstrap corrects for this overestimation, allowing for the correct inference of the Kappa statistic
- Unrealistic that the true level of within-category heterogeneity will be known for real world data
  - Bias in the analytic variance of Kappa is largely driven by the presence of this heterogeneity
  - Application of our bootstrap variance estimate does not harm inference in settings where no heterogeneity is present
  - Adoption of our methodology will provide robust inference of the Kappa statistic

Conclusions

- Even with slight amounts of variability among classification probabilities within categories, Kappa's analytic variance largely overestimates the true variance
  - Application of the bootstrap corrects for this overestimation, allowing for the correct inference of the Kappa statistic
- Unrealistic that the true level of within-category heterogeneity will be known for real world data
  - Bias in the analytic variance of Kappa is largely driven by the presence of this heterogeneity
  - Application of our bootstrap variance estimate does not harm inference in settings where no heterogeneity is present
  - Adoption of our methodology will provide robust inference of the Kappa statistic
- Further results can be seen in **Ryan**, Spotnitz, & Gillen (2020) "Variance estimation for the Kappa statistic in the presence of clustered data and heterogeneous observations", *Statistics in Medicine*. doi.org/10.1002/sim.8522[12]

Future Directions: Group Sequential Design

- For study, surgeons were flown out to central testing/training site in two groups of 7
- Observed kappas were much higher than the 0.6 null - did we need all 14?

Future Directions: Group Sequential Design

- For study, surgeons were flown out to central testing/training site in two groups of 7
- Observed kappas were much higher than the 0.6 null - did we need all 14?
- Can we make this study more efficient using sequential sampling/group sequential design?

Future Directions: Group Sequential Design

- Study framework used to assess early signs of study futility or efficacy
- Hypothesis tests performed at multiple points throughout data accrual (**interim analyses**) to determine if sufficient evidence to draw a conclusion early

Future Directions: Group Sequential Design

- Study framework used to assess early signs of study futility or efficacy
- Hypothesis tests performed at multiple points throughout data accrual (**interim analyses**) to determine if sufficient evidence to draw a conclusion early

Future Directions: Group Sequential Design

- Performing maximum of J planned analyses
- Statistic of interest at analysis j $\hat{\theta}^{(j)}$, $j \in \{1, ..., J\}$
- **Continuation set** $C_j = (a_j, b_j] \cup [c_j, d_j)$, $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$
- **Stopping set** $S_j \equiv C_j^c$
- At final analysis J:
  - $a_J = b_J = c_J = d_J$
- Think of a_j, b_j, c_j, d_j as critical values (**stopping boundaries**)
  - For one-sided $(\theta > \theta_0)$ test:
    - $\hat{\theta}^{(j)} \leq a_j$: stop study in favor of null (futility)
    - $\hat{\theta}^{(j)} \geq d_j$: stop study in favor of alternative (efficacy)
    - $\hat{\theta}^{(j)} \in (a_j = b_j = c_j, d_j)$: continue to analysis $(j+1)$
  - Need to adjust critical values we compare our statistic to at each analysis in order to maintain type I error
    - Need to know sequential pdf find appropriate values

Future Directions: Group Sequential Design

- Assume θ is Normally distributed
- Independent increments property:

$$Cov[\theta^{(j)}, \theta^{(j')}] = Var[\theta^{(j)}] = \sigma^{2,(j)}, j < j'$$

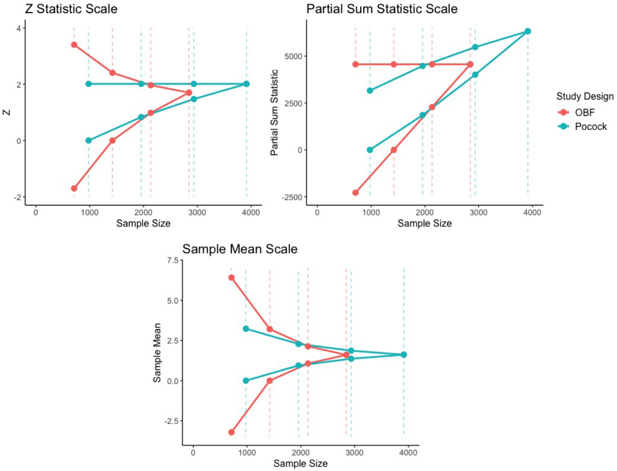
- Armitage et al.[1] and Jennison & Turnbull[7] found that, given Normal approximation of independent increments, **probability density function**, $\theta^{(j)}$, can be written as recursive Normal distributions:

$$f_j(\theta^{(j)}) = \begin{cases} \int_{C_{j-1}} f_{j-1}(u) \frac{1}{\sqrt{2\pi\sigma^{2,(j)}}} \exp\{-\frac{1}{2\sigma^{2,(j)}}(\theta^{(j)} - u)^2\} du, & \theta^{(j)} \notin C_{j-1} \\ 0, & \text{otherwise} \end{cases}$$

  - $f_{j-1}(u)$: density at previous analysis *(j-1)*
  - C_{j-1}: Continuation set for analysis *(j-1)*

Future Directions: Group Sequential Design

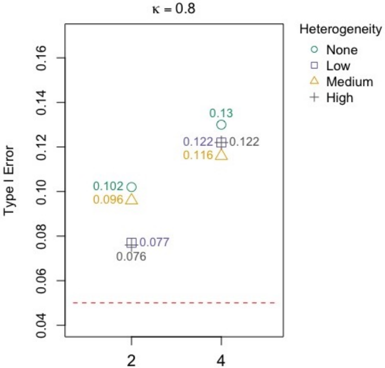
- Infinite number of combinations of a_j, b_j, c_j, d_j that will given us correct type I error (use sequential pdf to check)
    - Similar combinations with certain properties get grouped together and called **boundary shapes**
- Common boundary shapes:
    - **Pocock**[11]
    - **O'Brien-Fleming**[10]
        - More conservative earlier in the study



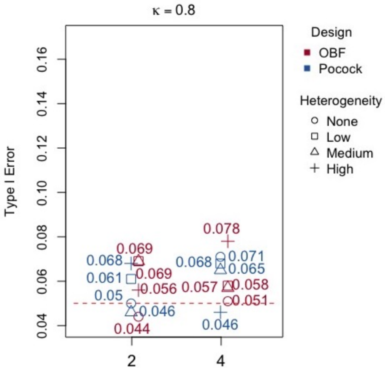
Future Directions: Group Sequential Design

- Issue: Kappa doesn't have independent increments property ⇒ difficult to find the sequential pdf to determine correct stopping boundaries

Naive Boundaries



Group Sequential Boundaries



- Using traditional GSD boundaries assuming independent increments doesn't quite control type I error, even if using bootstrapped variance

Future Directions: Group Sequential Design

- One solution: can use regular GSD boundaries for first (J-1) analyses, then simulate the last boundary necessary to maintain type I error

Future Directions: Group Sequential Design

- One solution: can use regular GSD boundaries for first (J-1) analyses, then simulate the last boundary necessary to maintain type I error
    - Not much help if you aren't making it to the final analysis
    - If never making it to final analysis, must be underestimating variance (smaller variance $\Rightarrow$ larger Z test statistic)

Future Directions: Group Sequential Design

- One solution: can use regular GSD boundaries for first (J-1) analyses, then simulate the last boundary necessary to maintain type I error
  - Not much help if you aren't making it to the final analysis
  - If never making it to final analysis, must be underestimating variance (smaller variance $\Rightarrow$ larger Z test statistic)
  - A way to rescale the variance?

Future Directions: Group Sequential Design

- Something that seems to be working:

Algorithm 2: GSD bootstrap algorithm for variance of Kappa statistic.

```
for j in J do
    if j==1 then
        Perform Algorithm 1 to obtain $\hat{\sigma}_B^{2,(1)}$ for $n_1$ surgeons;
    else
        Bootstrap $\kappa_b^{(1)}$ as in Algorithm 1 for $n_1$ surgeons;
        for u in 2:j do
            Bootstrap $\kappa_b^u$ as in Algorithm 1 for $n_u - n_{(u-1)}$ surgeons;
            Create $\kappa_b^{(u)}$ using bootstrapped $\sum_{v=1}^u n_v$ surgeons;
            $z_b^{(u)} = \frac{\kappa_b^{(u)} - \kappa_0}{(u-1)\hat{\sigma}_B^{2,(u)}/u}$;
            Compare $z_b^{(u)}$ to stopping boundary for analysis $u$ - if crosses,
              filter out all $z_b^{(u+1)}$, ... and $\kappa_b^{(u+1)}$, ...;
        end
        Calculate $\bar{\varphi}^{(u)} = \frac{1}{B}\sum_{b=1}^B \varphi_b^{(u)}$;
        Calculate $\hat{\sigma}_B^{2,(j)} = \frac{\sum_{b=1}^B \varphi_b^{(u)} - \bar{\varphi}^{(u)}}{B-1}$
    end
    Use $\frac{(j-1)}{j}\hat{\sigma}_B^{2,(j)}$ in Z-statistic to compare to stopping boundaries;
end
```

Acknowledgements

Thank you to Biom'up, SA, for the use of their SPOT GRADE study data.

email: marymr@uci.edu

Slides available at: maryryan.netlify.app

References I

[1] P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969.

[2] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960.

[3] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. Place: US Publisher: American Psychological Association.

[4] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. Place: US Publisher: American Psychological Association.

[5] J. L. Fleiss, J. Cohen, and B. S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.

[6] R. Gonin, S. R. Lipsitz, G. M. Fitzmaurice, and G. Molenberghs. Regression modelling of weighted by using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1):1–18, 2000. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9876.00175.

References II

[7] C. Jennison and B. W. Turnbull. Group-Sequential Analysis Incorporating Covariate Information. *Journal of the American Statistical Association*, 92(440):1330–1341, Dec. 1997.

[8] C. Kang, B. Qaqish, J. Monaco, S. L. Sheridan, and J. Cai. Kappa statistic for clustered dichotomous responses from physicians and patients. *Statistics in Medicine*, 32(21):3700–3719, 2013.

[9] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.

[10] P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, Sept. 1979.

[11] S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, Aug. 1977.

[12] M. M. Ryan, W. D. Spotnitz, and D. L. Gillen. Variance estimation for the Kappa statistic in the presence of clustered data and heterogeneous observations. *Statistics in Medicine*, 11(1), Jan. 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8522.

References III

[13] W. D. Spotnitz, D. Zielske, V. Centis, R. Hoffman, D. L. Gillen, C. Wittmann, V. Guyot, D. M. Campos, P. Forest, A. Pearson, and P. C. McAfee. The SPOT GRADE: A New Method for Reproducibly Quantifying Surgical Wound Bleeding. *Spine*, 43(11):E664, June 2018.

[14] Z. Yang and M. Zhou. Kappa statistic for clustered matched-pair data. *Statistics in Medicine*, 33(15):2612–2633, 2014.

[15] Z. Yang and M. Zhou. Kappa statistic for clustered physician–patients polytomous data. *Computational Statistics & Data Analysis*, 87:1–17, July 2015.

[16] M. Zhou and Z. Yang. A note on the kappa statistic for clustered dichotomous data. *Statistics in Medicine*, 33(14):2425–2448, 2014.