Stratified Micro-randomized Trials with Applications in Mobile Health

Abstract: Technological advancements in the field of mobile devices and wearable sensors make it possible to deliver treatments anytime and anywhere to users like you and me. Increasingly the delivery of these treatments is triggered by detections/predictions of vulnerability and receptivity. These observations are likely to have been impacted by prior treatments. Furthermore the treatments are often designed to have an impact on users over a span of time during which subsequent treatments may be provided. Here we discuss our work on the design of a mobile health smoking cessation study in which the above two challenges arose. This work involves the use of multiple online data analysis algorithms. Online algorithms are used in the detection, for example, of physiological stress. Other algorithms are used to forecast at each vulnerable time, the remaining number of vulnerable times in the day. These algorithms are then inputs into a randomization algorithm that ensures that each user is randomized to each treatment an appropriate number of times per day. We develop the stratified micro-randomized trial which involves not only the randomization algorithm but a precise statement of the meaning of the treatment effects and the primary scientific hypotheses along with primary analyses and sample size calculations. Considerations of causal inference and potential causal bias...
incurred by inappropriate data analyses play a large role throughout.
Outline

- Introduction to Mobile Health
- Sense²STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator
mHealth: Much Promise for Health Behavior Change!

- **Eating disorders** (e.g., Bauer et al., 2010)
- **Smoking cessation** (e.g., MD2K, 2017)
- **Physical activity** (e.g., Thomas & Bond, 2015)
- **Other chronic disorders** (e.g., Kristjánssóttir et al., 2013)
- **Alcohol use disorders** (e.g., Gustafson et al., 2014)
- **Mental illnesses** (e.g., Ben-Zeev et al., 2013)
Pushes are designed to help person over a particular time interval
Outline

- Introduction to Mobile Health
- Sense²STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator
This is a simplified version of Sense2Stop!

Stress State characterized by a combination of arousal and displeasure

(Kristensen, 1996; Posner et al., 2005)

Can be triggered by various circumstances in real-life

Shiffman et al. (2000) defined lapse as any occasion of smoking, even if only a puff. This was differentiated from relapse, which was defined as smoking at least five cigarettes for three consecutive days

The smoking cessation outcome can be 7-day point prevalence abstinence.

A “relapse” is defined as seven consecutive days of at least one puff per day following a period of total abstinence.
A “lapse” is defined as an isolated smoking episode of not more than six consecutive days followed by at least 24 h of abstinence.

This is from Ramelson, H. Z., Friedman, R. H., & Ockene, J. K. (1999). An automated telephone-based smoking cessation education and counseling system. Patient
Education and Counseling, 36(2), 131-144.
This is a prototype study as chest band is not for real life use. Also participants are paid a good bit.

**cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment**
Karen Hovsepian_, Mustafa al'Absiy, Emre Ertin, Thomas Kamarck^ Motohiro Nakajimay, Santosh Kumar at UbiComp’15, September 7-11, 2015,


Harris (2009) explains that cognitive defusion means:
Looking at thoughts rather than from thoughts
Noticing thoughts rather than becoming caught up in thoughts
Letting thoughts come and go rather than holding onto them
The general purpose of cognitive defusion is to:
Notice the true nature of thoughts – they are words or images in your mind
Respond to thoughts in terms of taking workable action – take action based on what “works” rather than what is “true”
Notice the actual process of thinking – recognize that thoughts do not dictate behaviors
Use cognitive defusion when thoughts are acting as a barrier to living in accordance with your true values

Cognitive restructuring (CR) is a psychotherapeutic process of learning to identify and dispute irrational or maladaptive thoughts known as cognitive distortions,[1] such as all-or-nothing thinking (splitting), magical thinking, over-generalization, magnification,[1] and emotional reasoning, which are commonly associated with many mental health disorders.[2] CR employs many strategies, such as Socratic questioning, thought recording, and guided imagery, and is used in many types of therapies, including cognitive behavioral therapy (CBT) and rational emotive therapy (RET). A number of studies demonstrate considerable efficacy in using CR-based
therapies.[3][4][5]
This is a simplified version of Sense2Stop!

State characterized by a combination of arousal and displeasure (Kristensen, 1996; Posner et al., 2005)
Can be triggered by various circumstances in real-life

Shiffman et al. (2000) defined lapse as any occasion of smoking, even if only a puff. This was differentiated from relapse, which was defined as smoking at least five cigarettes for three consecutive days

The smoking cessation outcome can be 7-day point prevalence abstinence.

It has been described that up to 80% of quitters eventually relapse within 6 months. A “relapse” is defined as seven consecutive days of at least one puff per day following a period of total abstinence.
A “lapse” is defined as an isolated smoking episode of not more than six consecutive days followed by at least 24 h of abstinence.
This is from Ramelson, H. Z., Friedman, R. H., & Ockene, J. K. (1999). An automated
Observations at different time scales

Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity. Hence, a total of 42% data is expected to be available from 16 hours of sensor wearing per day. Given these expectation, can we come up with a threshold for missing data from stress assessment?

• Primary: Should the smartphone notify the user with a reminder to utilize app directed stress-management exercises when the user is (not) stressed?
  – Does this effect vary with time or with current context?
• In the near term the reminder notification should reduce:
  – Near time, proximal, stress
Stratified Micro-Randomized Trial

- On each participant: \( O_1, A_1, \ldots, O_t, A_t, \ldots \)

- \( t \) : Decision time (times at which a treatment might be provided)

- \( O_t \): observations after time \( t - 1 \) and up to and including time \( t \)
  - \( X_t \subset O_t \) : time-varying stratification variable

Delta will be 120
Stratified Micro-Randomized Trial

• On each participant: $O_1, A_1, \ldots, O_t, A_t, \ldots$

• $t$: Decision time (times at which a treatment might be provided)
  • Sense$^2$STOP: each minute

• $O_t$: observations after time $t-1$ and up to and including time $t$
  • $X_t \subseteq O_t$: time-varying stratification variable
  • Sense$^2$STOP: $X_t = 1$ each minute if classified as stressed and $=0$, otherwise
Stratified Micro-Randomized Trial

- $O_t$: observations after time $t - 1$ and up to and including time $t$
- $X_t \subseteq O_t$: time-varying stratification variable
- Sense$^2$STOP: $X_t = 1$ each minute if classified as stressed and $=0$, otherwise

- $I_t \subseteq O_t$: availability indicator
- Sense$^2$STOP: $I_t = 1$ if not treated in prior hour and if online classification is possible; $=0$ otherwise

Available if data is good, if $t$ is at top of episode, if not driving, if no ema in prior 10 min, no treatment in prior hour.
Stratified Micro-Randomized Trial

- $O_t$: observations after time $t - 1$ and up to and including time $t$
  - $I_t \subseteq O_t$: availability indicator
  - Sense$^2$STOP: $I_t = 1$ if not treated in prior hour and if online classification is possible; $=0$ otherwise

- $A_t$: Randomized Treatment at decision time $t$
  - Sense$^2$STOP: $A_t = 1$ if reminder is pushed to participant; $=0$ otherwise
Delta will be 120 but in the sizing of the study we were thinking 60 min.
Why Micro-Randomize?

- In randomization ensures that we can assess causal effects of the reminder.
  - Should the smartphone remind you to practice a stress-regulation exercise when (not) stressed?
  - Does this effect vary with time and/or current context?
- Sequential randomization due to sequences of treatment

We don’t want to send these reminders if you don’t need them!
From Peng

Originally there are 64 person day satisfying at least 12 hours. I did not consider the 10 the person-day in which there is no episode classified as Stress, thus resulting in 54 person days (I did this because the KL divergence is undefined in these cases.)

Below is the summary of stress and non-stress episodes.

(1) the entire 64 person-days

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stress</td>
<td>0.000</td>
<td>2.000</td>
<td>4.000</td>
<td>4.828</td>
<td>7.000</td>
<td>17.00</td>
</tr>
<tr>
<td>Not Stress</td>
<td>3.00</td>
<td>19.00</td>
<td>27.50</td>
<td>28.55</td>
<td>37.50</td>
<td>55.00</td>
</tr>
</tbody>
</table>

(2) after removing 10 person-days with no Stress

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stress</td>
<td>1.000</td>
<td>3.000</td>
<td>5.000</td>
<td>5.722</td>
<td>7.750</td>
<td>17.00</td>
</tr>
<tr>
<td>Not Stress</td>
<td>3.00</td>
<td>22.25</td>
<td>29.00</td>
<td>29.98</td>
<td>39.00</td>
<td>55.00</td>
</tr>
</tbody>
</table>

---

Why Stratify?

- Fraction of decision points in one strata is low compared to other strata
  - Stratify the randomization to ensure sufficient treatment/no treatment in each strata.

- Sense²STOP:
  - On average 1 minute stressed for each 5.3 minutes not stressed (+ high within and between user variability)
From our paper, each stress episode lasts around 12 min and each non-stress episode lasts 11 min.
Stratified Micro-Randomized Trial

- Generally there is a “soft” budget, $\bar{N}_x$, for the number of treatments that can be provided over $T$ decision times for each strata of decision times.
  - $E\left[\sum_{t=1}^{T} A_t 1_{X_t=x} I_t\right] \approx \bar{N}_x$
  - Budget is usually due to participant burden concerns

- The budget is $E\left[\sum_{t\in day} A_t 1_{X_t=x} I_t\right] \approx 1.5$, for $x=0, 1$

Delta will be 120
Randomization probabilities

- Given $H_t = \{O_1, A_1, ..., O_t\}$, $X_t = x$ and $I_t = 1$, we deliver the treatment at time $t$ with probability

$$p_t(H_t) = \frac{\bar{N}_x - \sum_{s=1}^{t-1} C_{t, \lambda}(x)}{1 + g(x | H_t)}$$

- $\bar{N}_x$ is desired average no. of treatments
- $C_{t, \lambda}(x)$: soft version of the number of treatments that have already been delivered that day
- $g(x | H_t)$: forecast of number of available decision points at level $x$ remaining during day
Observations at different time scales

Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity. Hence, a total of 42% data is expected to be available from 16 hours of sensor wearing per day.
Outline

- Introduction to Mobile Health
- Sense²STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator
Observations at different time scales

Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity. Hence, a total of 42% data is expected to be available from 16 hours of sensor wearing per day. Given these expectation, can we come up with a threshold for missing data from stress assessment?

• Primary: Should the smartphone notify the user with a reminder to utilize app directed stress-management exercises when the user is (not) stressed?
  – Does this effect vary with time or with current context?
• In the near term the reminder notification should reduce:
  – Near time, proximal, stress
To make English precise use potential outcomes

- $\tilde{A}_t = \{A_1, ..., A_t\}$ (random treatments)
- $\tilde{a}_t = \{a_1, ..., a_t\}$ (realizations of treatments)
- $Y_{t,\Delta}(\tilde{a}_{t+\Delta-1})$ (one potential proximal response)
- $I_t(\tilde{a}_{t-1})$ (one potential availability indicator)
- $H_t(\tilde{a}_{t-1})$ (one potential history vector)
Treatment effect

Define the individual level effect as a contrast between two excursions:

\[
Y_{t,\Delta}(\bar{A}_{t-1,1}, a_{t+1} = 0, \ldots, a_{t+\Delta-1} = 0) - Y_{t,\Delta}(\bar{A}_{t-1,0}, a_{t+1} = 0, \ldots, a_{t+\Delta-1} = 0)
\]
Treatment effect

Define the individual level effect as a contrast between two excursions:

\[ Y_{t,\Delta}(\bar{A}_{t-1}, 1, a_{t+1} = 0, ..., a_{t+\Delta-1} = 0) \]
\[ -Y_{t,\Delta}(\bar{A}_{t-1}, 0, a_{t+1} = 0, ..., a_{t+\Delta-1} = 0) \]

Causal
Because we are designing a study, we attempt to impose minimal assumptions. This is so we don’t end up accidentally constraining the variety of analyses other scientists would like to do.

\[
Y_{t,\Delta}(\tilde{A}_{t-1}, 1, a_{t+1} = 0, ..., a_{t+\Delta-1} = 0) - Y_{t,\Delta}(\tilde{A}_{t-1}, 0, a_{t+1} = 0, ..., a_{t+\Delta-1} = 0)
\]

Absent assumptions above is not estimable
Treatment effect

$$Y_{t,\Delta}(\bar{A}_{t-1}, 1, a_{t+1} = 0, ..., a_{t+\Delta-1} = 0)$$

$$- Y_{t,\Delta}(\bar{A}_{t-1}, 0, a_{t+1} = 0, ..., a_{t+\Delta-1} = 0)$$

Define $\beta(t; x) = E[(Y_{t,\Delta}(\bar{A}_{t-1}, 1, 0) - Y_{t,\Delta}(\bar{A}_{t-1}, 0, 0))|I(\bar{A}_{t-1}) = 1, X_t(\bar{A}_{t-1}) = x]$.

$\beta(t; x)$: causal excursion effect at time $t$ beginning in strata $x$.
Marginal, Conditional & Causal

$\beta(t; x)$: effect at time $t$ in strata $x$

$E[(Y_{t,\Delta}(\tilde{A}_{t-1}, 1, \tilde{a}) - Y_{t,\Delta}(\tilde{A}_{t-1}, 0, \tilde{a}))I(\tilde{A}_{t-1}) = 1, X_t(\tilde{A}_{t-1}) = x]$

- Expectation is over the distribution of the potential outcomes
- Effect is marginal over past, $\tilde{H}_t(\tilde{a}_{t-1})$
- Effect is conditional on availability and in strata $x$ at time $t$. 
Expression for Treatment Effect $\beta(t; x)$

- The sequential randomization + randomization probabilities bounded away from 0, 1 imply that $\beta(t; x)$ can be expressed in terms of the data distribution:

$$E \left[ \prod_{j=t+1}^{t+\Delta-1} \frac{1[A_j = 0]}{p_j(H_j)^{A_j} (1 - p_j(H_j))^{1-A_j}} Y_{t,\Delta} | A_{t} = 1, H_{t} \right] | t_{t} = 1, X_{t} = x$$

$$-E \left[ \prod_{j=t+1}^{t+\Delta-1} \frac{1[A_j = 0]}{p_j(H_j)^{A_j} (1 - p_j(H_j))^{1-A_j}} Y_{t,\Delta} | A_{t} = 0, H_{t} \right] | t_{t} = 1, X_{t} = x$$

- $p_j(H_j)$ is randomization probability
Outline

- Introduction to Mobile Health
- Sense²STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator
Primary Hypothesis

- Consider decision points at which the individual is classified as stressed.
- We aim to contrast two treatment “excursions:”
  - (A) treatment now, no further treatment over subsequent 1 hour versus
  - (B) no treatment now, no further treatment over subsequent 1 hour
- Proximal response is fraction of time stressed over subsequent 1 hour.

Very imprecise english
Test Statistic

• Test statistic to test

\[ H_0: \{ \beta(t; x) \}_{t=1,\ldots,T, x=0,1} = 0 \]

(e.g. is there anything going on here?!) 

• Construct test statistic to target particular alternatives; consider alternatives of the form:
  • \( \beta(t; x) = f_t(x)' \beta \) where \( f_t(x) \in \mathbb{R}^q \) is feature vector depending on \( t \) and \( x \)
Test Statistic

- Primary Hypothesis
  \[ H_0: \{ \beta(t; x) \}_{t=1,...,T, x=0,1} = 0 \]

- Focus on alternatives of the form:
  - \( \beta(t; x) = f_t(x)' \beta \) where \( f_t(x) \in \mathbb{R}^q \) is feature vector depending on \( t \) and \( x \)
  - If we suspect a decreasing effect with day in study, we might try to capture this in a coarse manner:
    - \( f_t(x)' = (x, xd_t, xd_t^2, (1 - x), (1 - x)d_t, (1 - x)d_t^2) \) or
    - \( f_t(x)' = (x, xd_t, (1 - x), (1 - x)d_t) \)

\( d_t \) is day in study at decision point \( t \)

Explain what you do if you want to just focus on \( x=1 \), stressed times.
Control variables

Used to reduce the variance/increase power

- Control variables will be used in a working model for the average proximal response:

\[
E(0.5Y_{t,\Delta}(\bar{A}_{t-1}, 1, \bar{0}) + 0.5Y_{t,\Delta}(\bar{A}_{t-1}, 0, \bar{0})|H_t) = g_t(H_t)'\alpha
\]

for \( g_t(H_t) \), a vector of summaries of prior data.

- The test statistic/Type 1 error rate will be robust to the mis-specification of this working model.
Weighted-centered least squares criteria

To construct the test statistic calculate

$$\arg \min_{\alpha, \beta} P_n \left[ \sum_{t=1}^{T} I_t w_t(H_{t+\Delta-1})(Y_{t,\Delta} - g_t(H_t)' \alpha - (A_t - .5)f_t(X_t)' \beta)^2 \right]$$

- $P_n$ means average over participants’ data
- $w_t(H_{t+\Delta-1}) = \frac{\prod_{s=1}^{\Delta-1} 1[A_{t+s}=0]}{\prod_{s=0}^{\Delta-2} p_{t+s} A_{t+s} (1-p_{t+s})^{1-A_{t+s}}}$
- $p_{t+s}$ is the randomization probability used in the study

Explain why $A_t-0.5$?!
Weighted-centered least squares criteria

To construct the test statistic calculate

$$\arg \min_{a,\beta} P_t \left[ \sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1}) (y_{t,\Delta} - g_t(H_t)' \alpha - (A_t - .5) f_t(x_t)' \beta)^2 \right]$$

This results in $\hat{\beta}$. 
An Aside

\( \hat{\beta} \) is an estimator of

\[
\beta^* = \arg \min_\beta E \left[ \sum_{t=1}^{T} I_t(\beta(t; X_t) - f_t(X_t)'\beta)^2 \right]
\]

Recall \( f_t(x) \in \mathbb{R}^q \) is feature vector depending on \( t \) and \( x \)
Test Statistic

To construct the test statistic calculate

$$\hat{\beta} = \arg \min_{\alpha, \beta} \sum_{t=1}^{T} l_{w_t(H_{t+\Delta-1})}(Y_{t,\Delta} - g_t(H_t)'\alpha - (A_t - .5)f_t(X_t)'\beta)^2$$

This results in $\hat{\beta}$.

We also construct an estimator of the standard error of $\sqrt{n}\hat{\beta}$ ($n$ is the sample size): $\hat{\Sigma}$

This standard error must allow for unspecified correlation across time in the $Y_{t,\Delta}$.
Test Statistic

\[ T_n = n \hat{\beta}' \hat{\Sigma}^{-1} \hat{\beta} \]

We also construct an estimator of the standard error of \( \sqrt{n} \hat{\beta} \) (\( n \) is the sample size):

\[ \hat{\Sigma} \]

This standard error allows for unspecified correlation across time in the \( Y_{t,\Delta}'s \)

\( \hat{\Sigma} \) is a robust standard error -- sandwich formula
Use two different small sample corrections, one in test statistic and the other in the critical value
Outline

- Introduction to Mobile Health
- Sense²Stop and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator
Sample size formula

- Define $\gamma = (\beta^*)\Sigma^{-1}(\beta^*)$

- Then the sample size is the smallest value $n$ such that

$$1 - F_{q,n-(q'+q)n\tau} \left( \frac{n - (q' + 1)}{n - (q' + q)} \cdot F_{q,n-(q'+q)\sigma_0}^{-1}(1 - \alpha_0) \right) \geq 1 - \beta_0$$

where $\alpha_0$ is the Type I error rate and $1 - \beta_0$ is the power.
Inputs into the sample size calculation

- Desired Type I and Type II error rates,
- Targeted alternative \( \{ \beta(t; x) \}_{t=1, \ldots, T, x=0,1} \),
- Selected “control variables” \( g_t(H_t) \),
- The randomization formula used to determine \( P(A_t = 1 \mid h_t) \) given any history, \( H_t = h_t \) and
- A generative model for \( \{ H_t \}_{t=1, \ldots, T} \).
“Data-driven” baseline generative model

We construct summary statistics using subset of the data collected in an observational, no treatment, smoking cessation study of 50 cigarette smokers

- For each episode type (i.e., \( x \in \{0,1\} \)), estimate the probability that the next episode will be a stress episode – i.e., a 2 by 1 vector \( W \)
- For each episode type (i.e., \( x \in \{0,1\} \)), estimate the average episode length–i.e., a 2 by 1 vector \( Z \)
“Data-driven” baseline generative model

We model the joint process \((X_t, U_t)\) by a Markov chain
- \(X_t\): Episode classification
- \(U_t\): Stage of episode

Inputs:
- \(\bar{W} = (0.67, 0.519)\) (probability next episode is “stress”)
- \(\bar{Z} = (10.9, 12.0)\) (average duration of episode in min.)

Two values for \(\bar{W}, \bar{Z}\) correspond to not-stress, stress starting state
Generative model under treatment

• \((X_t, U_t)\), is a Markov chain
  • \(X_t\) : Episode classification
  • \(U_t\) : Stage of episode

• We compute the alternative transition matrix that achieves the targeted alternative. This becomes the transition matrix for 60 time points after a treatment.

Type I=.05
Type II=.2
Sample Size Calculation

- Type I error = 0.05; Type II error = 0.20
- Targeted Alternative: \( \beta(t; x) = f_t(x)\beta \)
- Working model for average proximal response \( g_t(H_t)\alpha \)
  - where \( f_t(x)' = g_t(x)' = (x, xd_t, xd_t^2, (1 - x), (1 - x)d_t, (1 - x)d_t^2) \)

Results:

Table 1: Estimated sample size, \( n \), and achieved power.

<table>
<thead>
<tr>
<th>( \beta )</th>
<th>Sample size</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.030</td>
<td>50</td>
<td>80.6</td>
</tr>
<tr>
<td>0.025</td>
<td>67</td>
<td>80.7</td>
</tr>
<tr>
<td>0.020</td>
<td>127</td>
<td>80.6</td>
</tr>
</tbody>
</table>

(1) an initial conditional effect\(= 0 \), (2) the day of maximal effect = 5 days and (3) the average conditional treatment effect is the same for both \( x=0 \) (not stress) and \( x=1 \) (stress)—these are the rows in table 1
Increasing Robustness to misspecification of the transition matrix

Recall \( (X_t, U_t) \), is a Markov chain

- \( X_t \): Episode classification
- \( U_t \): Stage of episode

We compute the largest sample size under the inputs:

- \( \bar{W} = (.067 \pm \epsilon , .519 \pm \epsilon) \)
- \( \bar{Z} = (10.9 \pm \epsilon', 12.0 \pm \epsilon') \)

<table>
<thead>
<tr>
<th>( \bar{W} )</th>
<th>Sample size</th>
<th>Minimum Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = 0.030 )</td>
<td>90</td>
<td>81.9</td>
</tr>
<tr>
<td>( \beta = 0.025 )</td>
<td>107</td>
<td>80.4</td>
</tr>
<tr>
<td>( \beta = 0.020 )</td>
<td>208</td>
<td>80.5</td>
</tr>
</tbody>
</table>

This is work in progress

For the average durations \( \bar{Z} \) we have:
Mean for NS, S: 10.9, 12.0 (in paper) with standard errors (0.12 for NS and 0.28 for S).
SD for NS, S: 6.89, 6.48

Choosing the new \( \bar{Z} \) should reflect sufficient deviation from the current \( \bar{Z} \),
but stay within the SDs of around 6.5 to 7.0. Thus the choice of 2

For \( \bar{W} \), I can calculate the SD for the estimated fractions (i.e., \( \bar{W} \)) obtained
from the data. This is 0.005 for NS and 0.03 for S. Given this the value 0.01 made
sense as potential deviations. I wanted to choose one level across both S and NS, and
thus opted for the two I chose.
Intuition Behind the Sample Sizes

- Because the alternative hypothesis targets a low dimensional alternative instead of all possible functions of time, $t$ and stress, $x$, the calculator is able to use within participant contrasts to increase power and ↓ sample size

- Sense$_2$STOP is a short study of only 10 days with few treatments per day, e.g. 3; this decreases power and ↑ sample size
- Small effect sizes decreases power and ↑ sample size
Micro-Randomized Trials: When are they (not) useful?

- **NOT USEFUL.** When circumstances are rare: Want to learn the best type of alert to prevent suicide attempt
- **USEFUL.** When circumstances change rapidly: Stress, urges to smoke, adherence, physical activity
- **NOT USEFUL.** Proximal outcome cannot be easily sensed or observed: suicide ideation, craving
- **USEFUL.** Proximal outcome can be unobtrusively sensed or unobtrusively self-reported.

Last two are in a state of flux and depend on the current sensor technology.
pinpointing where in stress episode to intervene
Language regarding what times constitute not stressed times
How to handle missing data
Tradeoff between sensitivity/specificity of stress classifications and the number of times during the day as which you can check if an intervention push is useful
Collaborators!