Safe Testing

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam
Mathematical Institute – Leiden University

joint work with Rianne de Heide, Wouter Koolen (and other group members to be involved too...)


Slate **Sep 10th 2016: yet another classic finding in psychology—that you can smile your way to happiness—just blew up…**

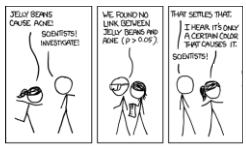
"at least 50% of highly cited results in medicine is irreproducible"
J. Ioannidis, PLoS Medicine 2005

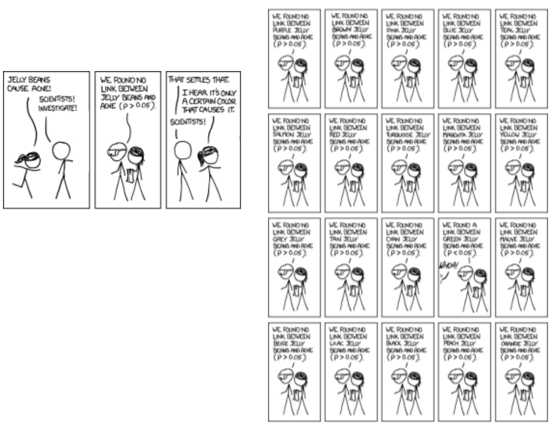
Reproducibility Crisis
Cover Story of Economist (2013), Wall Street Journal, Science (2012)

Reasons for Reproducibility Crisis

1. **Publication Bias**

2. Problems with Hypothesis Testing Methodology





Xkcd.org

Reasons for Reproducibility Crisis

1. Publication Bias

2. **Problems with Hypothesis Testing Methodology**

Reasons for Reproducibility Crisis

1. Publication Bias

2. **Problems with...**

p-values

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

80 years and still unresolved...

- Standard method for testing is still

  p-value-based
  null hypothesis significance testing

  ...an amalgam of Neyman-Pearson's and Fisher's 1930s methods
  - everybody in psychology and medical sciences (and even in A/B testing) does it...
  - .... most statisticians agree it's not o.k....
  - ...but still can't agree on what to do instead!

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, today we assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}$, $\Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, assume $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}$, $\Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$ **Simple H_0**

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  H_1 : $X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

  $H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$

  $H_1 = \{ P_{\sigma,\mu} | \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \}$

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. **Composite H_0**

  $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

  $H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$

  $H_1 = \{ P_{\sigma,\mu} | \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \}$

P-value Problem #1: Combining Independent Tests

- Suppose two different research groups tested the same new medication. How to combine their test results?
- **You can't multiply p-values!**
  - **This will (wildly) overestimate evidence against the null hypothesis!**
  - Different valid p-value combination methods exist (Fisher's; Stouffer's) but give different results
- **In "our" method evidences can be safely multiplied**

P-value Problem #2: Combining Dependent Tests

- Suppose reseach group A tests medication, gets 'almost significant' result.
- ...whence group B tries again on new data. How to combine their test results?
  - **Now Fisher's and Stouffer's method don't work anymore – need complicated methods!**
- **In "our" method, despite dependence, evidences can still be safely multiplied**

P-value Problem #2b: Extending Your Test

- Suppose reseach group A tests medication, gets 'almost significant' result.
- **Sometimes group A can't resist to test a few more subjects themselves...**
  - In a recent survey **55% of psychologists** admit to have succumbed to this practice [L. John et al., *Psychological Science*, 23(5), 2012]
- **In "our" method, despite dependence, evidences can still be safely multiplied**

P-value Problem #2b: Extending Your Test

- Suppose reseach group A tests medication, gets 'almost significant' result.
- **Sometimes group A can't resist to test a few more subjects themselves...**
  - A recent survey revealed that **55% of psychologists** have succumbed to this practice
- But isn't this just **cheating?**
  - **Not clear: what if you submit a paper and the *referee* asks you to test a couple more subjects? Should you refuse because it invalidates your p-values!?**

Menu

1. A problem with/limitation of with p-values
2. S-Values and Safe Tests
   - ...solves the stop/continue problem
   - gambling interpretation
3. **The New Work:** Safe Testing for Composite H_0

S-Values: General Definition

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - Assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- An **S-value** for sample size n is a function $S : \mathcal{X}^n \to \mathbb{R}_0^+$ such that for **all** $P_0 \in H_0$, we have

$$\mathbf{E}_{X^n \sim P_0} [S(X^n)] \leq 1$$

First Interpretation: p-values

- Proposition: Let S be an S-value. Then $S^{-1}(X^n)$ is a conservative p-value, i.e. p-value with **wiggle room**:
- for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- Proof: just Markov's inequality!

$$P\left(S(X^n) \geq \alpha^{-1}\right) \leq \frac{\mathbf{E}[S(X^n)]}{\alpha^{-1}} = \alpha$$

Safe Tests

- The Safe Test against H_0 at level α based on S-value S is defined as the test which rejects H_0 if $S(X^n) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- ....the safe test which rejects H_0 iff $S(X^n) \geq 20$, i.e. $S^{-1}(X^n) \leq 0.05$, has **Type-I Error** Bound of 0.05

Interpretation 1(b): Type-I Error

- The Safe Test against H_0 at level α based on S-value S is defined as the test which rejects H_0 if $S(X^n) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- ....the safe test which rejects H_0 iff $S(X^n) \geq 20$, i.e. $S^{-1}(X^n) \leq 0.05$, has **Type-I Error** Bound of 0.05

First Example

1. H_0 and H_1 are point hypotheses:

$$S(X^n) = \frac{p_1(X^n)}{p_0(X^n)}$$

...is an S-value.

First Example

1. H_0 and H_1 are point hypotheses:

$$S(X^n) = \frac{p_1(X^n)}{p_0(X^n)}$$

...is an S-value, since

$$\mathbf{E}_{X^n \sim P_0}\left[\frac{p_1(X^n)}{p_0(X^n)}\right] = \sum_{x^n \in \mathcal{X}^n} p_0(x^n) \cdot \frac{p_1(x^n)}{p_0(x^n)} = \sum_{x^n \in \mathcal{X}^n} p_1(x^n) = 1.$$

...can be extended to general stopping times τ, densities, Radon-Nikodym derivatives etc...

Safe Tests are Safe under optional continuation

- Suppose we observe data $(X_1, Y_1), (X_2, Y_2), \ldots$
  - Y_i: side information, independent of X_i's
- Let $S_1, S_2, \ldots, S_k$ be an arbitrarily large collection of (potentially "identical") S-values for sample sizes $n_1, n_2, \ldots, n_k$ respectively. Let $N_j := \sum_{i=1}^{j} n_i$
- We first evaluate S_1 on data $(X_1, \ldots, X_{n_1})$.

Safe Tests are Safe under optional continuation

- Suppose we observe data $(X_1, Y_1), (X_2, Y_2), \ldots$
  - Y_i: side information, independent of X_i's
- Let $S_1, S_2, \ldots, S_k$ be an arbitrarily large collection of (potentially "identical") S-values for sample sizes $n_1, n_2, \ldots, n_k$ respectively. Let $N_j := \sum_{i=1}^{j} n_i$
- We first evaluate S_1 on data $(X_1, \ldots, X_{n_1})$.
- If outcome is in certain range (e.g. promising but not conclusive) and Y_{n_1} has certain values (e.g. 'boss has money to collect more data') then....
  we evaluate S_2 on data $(X_{n_1+1}, \ldots, X_{N_2})$,
  otherwise we **stop.**

Safe Tests are Safe

- We first evaluate S_1.
- If outcome is in certain range and Y_{n_1} has certain values then we evaluate S_2 on new batch of data; otherwise we **stop.**
- If S_2 is in certain range and Y_{N_2} has certain values then we perform S_3, else we **stop**.
- ...and so on

(note that sequentially computed S-values may but need not have identical definitions, but data must be different for each test!)

Safe Tests are Safe

- We first evaluate S_1.
- If outcome is in certain range and Y_{n_1} has certain values then we evaluate S_2 ; otherwise we **stop.**
- If outcome of S_2 is in certain range and Y_{N_2} has certain values then we compute S_3 , else we **stop**.
- ...and so on
- ...when we finally stop, after say K data batches, we report as final result the product $S := \prod_{j=1}^{K} S_j$
- **First Result, Informally: any S composed of S-values in this manner is itself an S-value, irrespective of the stop/continue rule used!**

Safe Tests are Safe

- We first evaluate S_1.
- If outcome is in certain range and Y_{n_1} has certain values then we evaluate S_2 ; otherwise we **stop.**
- If outcome of S_2 is in certain range and Y_{N_2} has certain values then we compute S_3 , else we **stop**.
- ...and so on
- ...when we finally stop, after say K data batches, we report as final result the product $S := \prod_{j=1}^{K} S_j$
- **First Result, Informally: any S composed of S-values in this manner is itself an S-value, irrespective of the stop/continue rule used!**

We solved a central problem of p-values!

Second, Main Interpretation: Gambling!

Safe Testing = Gambling!

Kelly (1956)

- At time 1 you can buy ticket 1 for 1\$. It pays off $S_1(X_1, \ldots, X_{n_1})$ \$ after n_1 steps
- At time 2 you can buy ticket 2 for 1\$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ \$ after n_2 further steps.... and so on.

You may buy multiple and fractional nrs of tickets.

Safe Testing = Gambling!

- At time 1 you can buy ticket 1 for 1\$. It pays off $S_1(X_1, \ldots, X_{n_1})$ \$ after n_1 steps
- At time 2 you can buy ticket 2 for 1\$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ \$ after n_2 further steps.... and so on.

You may buy multiple and fractional nrs of tickets.

- You start by investing 1\$ in ticket 1.

Safe Testing = Gambling!

- At time 1 you can buy ticket 1 for 1\$. It pays off $S_1(X_1, \ldots, X_{n_1})$ \$ after n_1 steps
- At time 2 you can buy ticket 2 for 1\$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ \$ after n_2 further steps.... and so on.

You may buy multiple and fractional nrs of tickets.

- You start by investing 1\$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2.

Safe Testing = Gambling!

- At time 1 you can buy ticket 1 for 1\$. It pays off $S_1(X_1, \ldots, X_{n_1})$ \$ after n_1 steps
- At time 2 you can buy ticket 2 for 1\$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ \$ after n_2 further steps.... and so on.

You may buy multiple and fractional nrs of tickets.

- You start by investing 1\$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on..

Safe Testing = Gambling!

- You start by investing 1\$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on...
- ***S* is simply your end capital**

Safe Testing = Gambling!

- You start by investing 1\$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on...
- ***S* is simply your end capital**
- Your don't expect to gain money, no matter what the stop/continuation rule since **none of individual gambles S_k are strictly favorable to you**

$$\mathbf{E}_{P_0}[S_1] \leq 1, \mathbf{E}_{P_0}[S_2] \leq 1, \ldots \Rightarrow \mathbf{E}_{P_0}[S] \leq 1$$

Safe Testing = Gambling!

- Hence a **large value of S** indicates that something has happened that is higly unlikely under H_0 ...
- **"Amount of evidence against H_0" is thus measured in terms of how much money you gain in a game that would allow you not to make money in the long run if H_0 were true!**
- **Optional Continuation is possible because "you don't expect to make money in a casino no matter what rule you use to decide when to go home"**

Menu

1. Some of the problems with p-values
2. Safe Testing with S-values
   - ...solves the optional continuation problem
   - gambling interpretation
3. **The New Work:** Composite H_0

Safe Testing and Bayes

- **Bayes factor hypothesis testing (Jeffreys '39)** with $H_0 = \{p_\theta | \theta \in \Theta_0\}$ vs $H_1 = \{p_\theta | \theta \in \Theta_1\}$:
  Evidence in favour of H_1 measured by

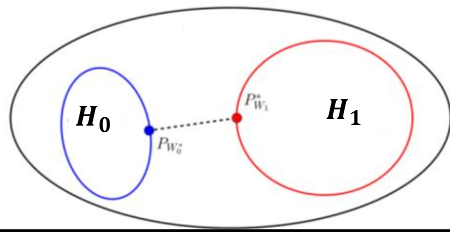
$$\frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$$

where

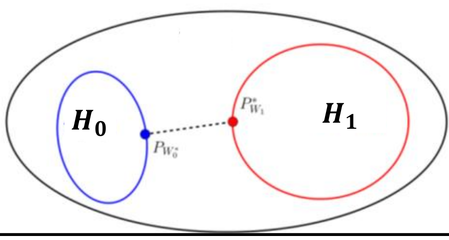
$$\bar{p}(X_1, \ldots, X_n \mid H_1) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) w_1(\theta) d\theta$$

$$\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$$

- In general, Bayes factors are not S-values
- But for some very special priors they always* are
- For every prior W_1^*, the prior W_0 achieving $\min_{W_0} D(P_{W_1^*} || P_{W_0})$ gives rise to an S-value
- D is KL divergence: W_0 is "(reverse) information projection"



- In general, Bayes factors are not S-values
- But for some very special priors they always* are
- For every prior W_1^*, the prior W_0 achieving $\min_{W_0} D(P_{W_1^*} || P_{W_0})$ gives rise to an S-value
- "best" S-value for (W_1^*, W_0^*) achieving $\min_{W_1} \min_{W_0} D(P_{W_1^*} || P_{W_0})$



Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing
between $H_0 = \{p_0\}$ and $H_1 = \{p_\theta | \theta \in \Theta_1\}$:
Evidence measured by

$$\frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$$

where

$$\bar{p}(X_1, \ldots, X_n \mid H_1) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) w_1(\theta) d\theta$$

$$\bar{p}(X_1, \ldots, X_n \mid H_0) := p_0(X_1, \ldots, X_n)$$

Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing
between $H_0 = \{ p_0 \}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

Take $S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{p_0(X_1, \ldots, X_n)}$

and note that (no matter what prior w_1 we chose)

$$\mathbf{E}_{X^n \sim P_0}[S(X^n)] =$$
$$\int p_0(x^n) \cdot \frac{\bar{p}(x^n \mid H_1)}{p_0(x^n)} dx^n = \int \bar{p}(x^n \mid H_1) dx^n = 1$$

Safe Testing and Bayes, simple H_0

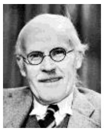
Bayes factor hypothesis testing
between $H_0 = \{ p_0 \}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

Take $S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{p_0(X_1, \ldots, X_n)}$

and note that (no matter what prior w_1 we chose)

$$\mathbf{E}_{X^n \sim P_0}[S(X^n)] = 1$$

The Bayes Factor for Simple H_0 is an S-value!



Composite H_0: Bayes may not be Safe!

Bayes factor given by $S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$

where $\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$

Composite H_0: Bayes may not be Safe!

Bayes factor given by $S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$

where $\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$

S-value requires that **for all** $P_0 \in H_0$:

$$\mathbf{E}_{X^n \sim P_0}[S(X^n)] \leq 1$$

...but for a Bayes factor we can only guarantee that

$$\mathbf{E}_{X^n \sim \bar{P}(\cdot | H_0)}[S(X^n)] \leq 1$$

Central Result: JIPR/RIPR (just a teaser…)

- For **completely arbitrary** composite H_1 and H_0 , one can construct nontrivial safe tests after all!
- These do take the form

$$S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$$

…after all, but for some very special priors on parameters on parameters in H_1 and H_0 (they are '**reverse** and **joint information projection** priors')
(these priors may be 'improper' (i.e. they do not integrate) and depend on sample size)

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$
σ^2 unknown ('nuisance') parameter
$H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$ $\quad H_1 = \{ P_{\sigma,\mu} | \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \}$

- In general Bayes factor tests are *not* safe
- But lo and behold, Jeffreys' uses very special priors and his Bayes factor is an S-value, so his Bayesian t-test is a Safe Test! But not the 'best' safe t-test...

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

σ^2 unknown ('nuisance') parameter

$H_0 = \{ P_\sigma | \sigma \in (0, \infty)\}$ $\quad H_1 = \{ P_{\sigma,\mu} | \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\}\}$

- In general Bayes factor tests are *not* safe
- But lo and behold, Jeffreys' uses very special priors and his Bayes factor is an S-value, so his Bayesian t-test is a Safe Test! But not the 'best' safe t-test...

Experimental Results/Conclusion

- With the GROW safe t-test you need to reserve about 20% more data points to obtain the same power at the same effect size, compared to the standard t-test
- …but you are allowed to do *optional stopping*: stop as soon as $S \geq 20$!
- Then **on average** you need about the same amount of data as with the standard t-test
- I wonder: is there a good excuse *not* to use the Safe t-test?