Construct validity

- Making sure it matches the actual phenomena in the real world that you are trying to model
- The "reproducibility crisis" is fundamentally a problem of over-reporting spurious results

Use Beta-binomial distributions rather than Gaussians

- Binomial: probability of something being a 0 and a 1 between two possible outcomes
- The conjugate prior of a binomial distribution is a beta distribution. This means that when the prior distribution is a conjugate prior for a certain likelihood function (such as the binomial distribution), the resulting posterior distribution belongs to the same family as the prior. You observe a series of binary outcomes and estimate the probability on any individual trial that it will be a success or a failure
- The only thing you need to calculate posterior is count of observations and whether they were successful or not
- The underlying distribution is binomial. There is a certain probability of drawing a zero and a certain probability of drawing a one. What the beta distribution tracks is what my beliefs about the underlying probability are. Alpha: 0 :: Beta : 1

Ways to report

- Mean of your observation (like 0.7). This is publishing your *results*. **Report the mean of a single urn**
- Number of successes and failures (gives you the uncertainty associated with your observation). This is publishing your *data*. This is an easy way to include uncertainty of an observation in addition to the results of the observation. **Report all data for one experiment (choose an urn, give a number of trials and successes). This is standard since you don't publish null results. Only publish one result, but it's the one that you think is significant.**
- **Choose a subset of data to report (drop observations)** for a single urn. This is a pretty good cover story and people will understand what we're asking them to do. Reasonable to say: "your objective is to choose a set of data to publish in a scientific journal that you think will be interesting to the community." You can say something like "it doesn't matter if you're truthful or not and the only thing that matters is if you get published". More interesting to look at what they *can* report as our variable that we are manipulating.

Specifying reasonable alternative models

- Show that people actually follow the model. Here's the phenomena: why does information aggregation in science lead to a reproducibility crisis?

- Looking at the incentive structure of science, let's write down a model that captures that structure: how do people gather data, report it, and how do editors choose which reports to publish.
- Can't test on science itself, so we'll do a behavioral experiment where we run a bunch of variations and show that our model captures people's response patterns to those different circumstances.
- For Reporting Strategy #1 (above)
    - If a scientist's utility function is purely whether it will be published, no one will ever give the mean.
    - Always truthful? Always report mean
    - Maximize published? Always lie
    - Reasonable hypothesis? Somewhere in the middle. Then we show that this "somewhere in the middle" hypothesis is able to account for the type of exaggeration that I am describing.

Simulation

- Depending on the variation of the experiment
    - Change scientists' reporting utility function so they ascribe a different utility to each of the possible reports OR
    - Change the set of reports that they *could* issue given the data
- Show that tweaking these knobs changes what they try to publish -> changes what gets published -> changes what scientific consensus ends up looking like
- Make a single-generation experiment with jsPsych and Psiturk. Here is a demo: https://github.com/fredcallaway/heroku-experiment. Use PsyNet for multi-generation experiments.
- Goal of simulation: show how different settings (incentive structures, message structure, etc.) affect the end outcome of what gets published
- Assumption: "people choose to report information that is likely to get published" which *leads* them to behavior where they exaggerate
- `p( d | h )`
    - `h:` certain underlying condition, which is the data observed
    - `d:` what you report. The set of all possible values for d is the set of possible reports that a scientist could issue.
    - Utility function: the utility that a scientist ascribes to different reports. In standard RSA, this would be "how surprising" the report is to a listener who has a uniform prior.
    - Softmax over the utility function to obtain a noisy-rational speaker model

Evaluating the posteriors

- Assumption: assume a uniform prior over whether it'll be a success or not. Alpha and beta is 1 (assume a listener starts with that prior). Evaluate the KL

divergence between what they believe and what they believe after seeing my scientific result? That tells you how surprising the result is.
- Plot the results for different betas similarly to how Ted does it in his stereotypes notebook. Choose an utterance based on different beliefs.
    - He ran simulations forward where he had individuals and did generational information passage. Then, he showed that under certain conditions, stereotypes emerge. Under others, stereotypes did not emerge. This is a similarly structured experiment.

Findings:
- The most interesting finding from this paper would be to show how the simulation settings affect the frequency of false results being published