# UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

# Content

# Replication and hypothesis testing

**Greg Francis**

# Experimental Methods

- Suppose you hear about two sets of experiments that investigate phenomena A and B
- Which effect is more believable?

| | Effect A | Effect B |
|---|---|---|
| Number of experiments | 10 | 19 |
| Number of experiments that reject $H_0$ | 9 | 10 |
| Replication rate | 0.9 | 0.53 |

# Replication

- **Effect A** is Bem's (2011) *precognition* study that reported evidence of people's ability get information from the future
  - I do not know any scientist who believes this effect is real
- **Effect B** is from a meta-analysis of the *bystander effect*, where people tend to not help someone in need if others are around
  - I do not know any scientist who does not believe this is a real effect
- So why are we running experiments?

| | Effect A | Effect B |
|---|---|---|
| Number of experiments | 10 | 19 |
| Number of experiments that reject $H_0$ | 9 | 10 |
| Replication rate | 0.9 | 0.53 |

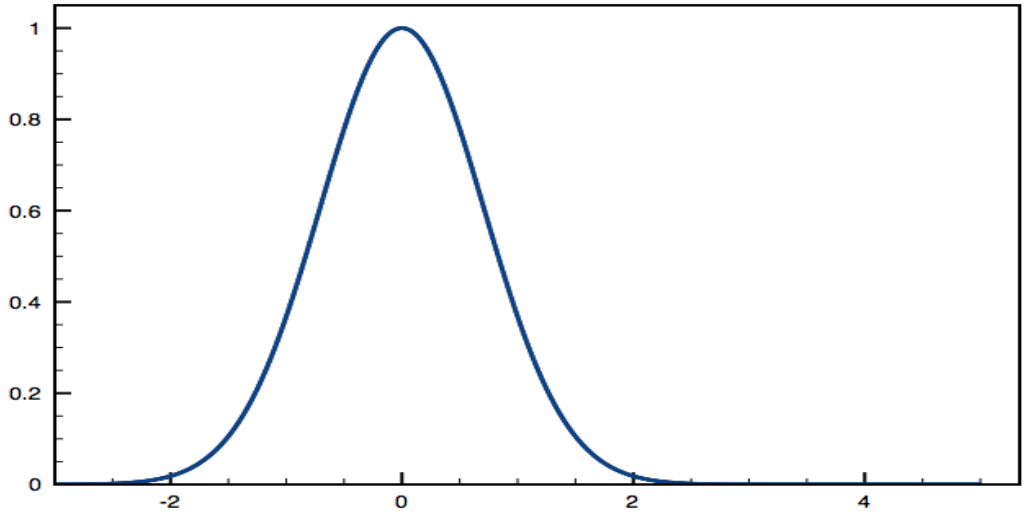
# Replication

- Replication has long been believed to be the final arbiter of phenomena in science
- But it seems to not work
  - Not sufficient (Bem, 2011)
  - Not necessary (bystander effect)
- In a field that depends on hypothesis testing, like experimental psychology, some effects should be rejected **because** they are so frequently replicated

# Hypothesis Testing (For Means)

- We start with a null hypothesis: no effect, $H_0$
- Identify a sampling distribution that describes variability in a test statistic

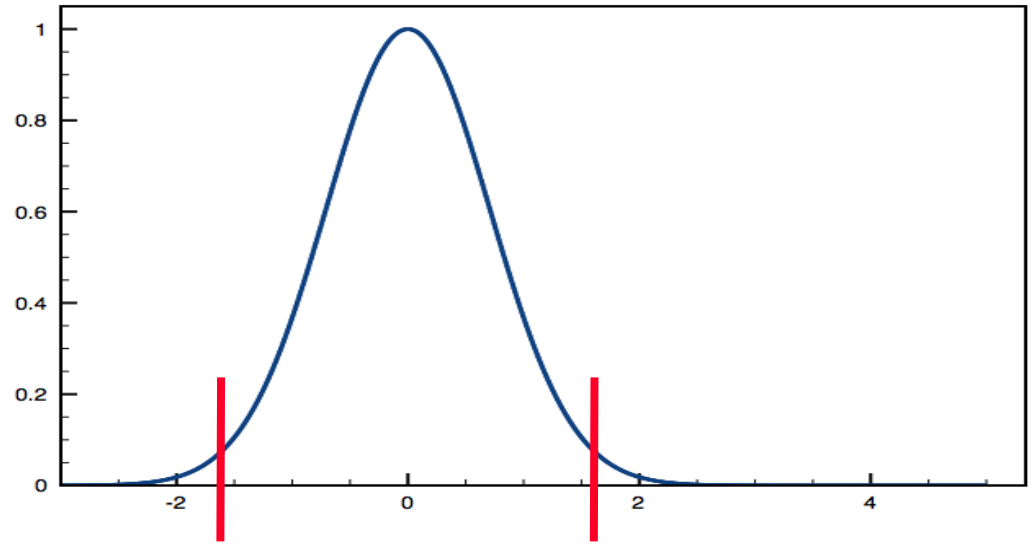
$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{\overline{X_1} - \overline{X_2}}}$$



# Hypothesis Testing (For Two Means)

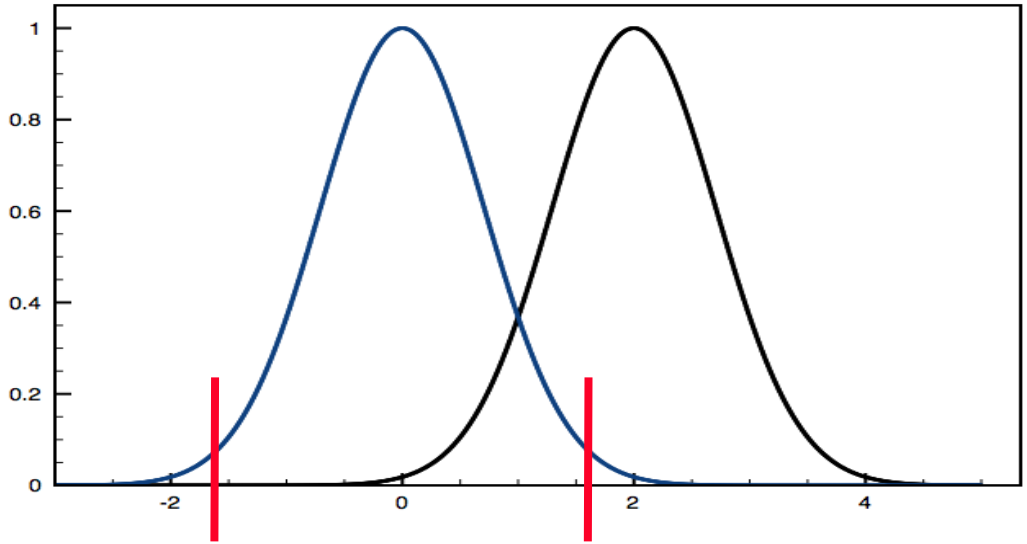
- We can identify rare test statistic values as those in the tail of the sampling distribution
- If we get a test statistic in either tail, we say it is so rare (usually 0.05) that we should consider the null hypothesis to be unlikely
- We reject the null

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{\overline{X_1} - \overline{X_2}}}$$



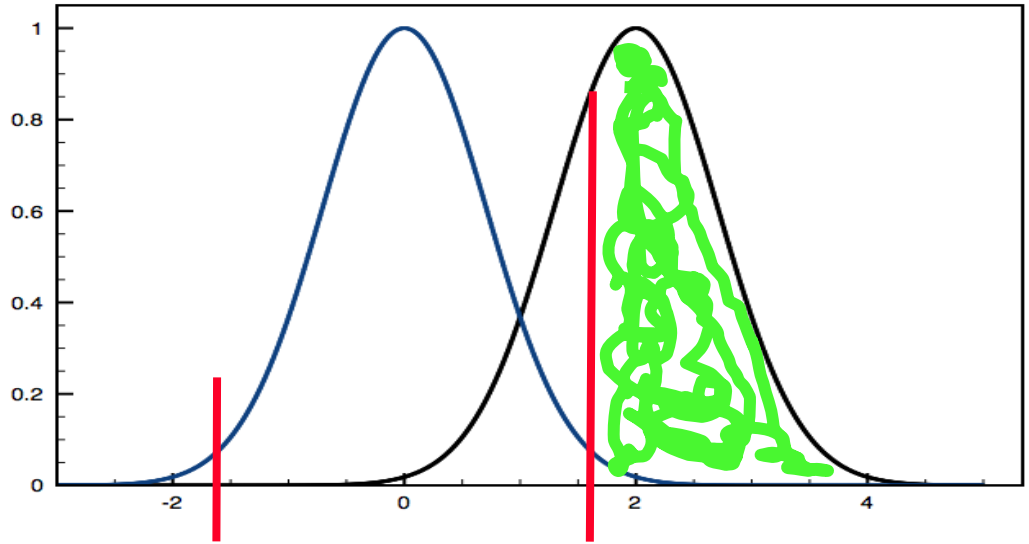
# Alternative Hypothesis

- If the null hypothesis is not true, then the data came from some other sampling distribution ($H_1$)



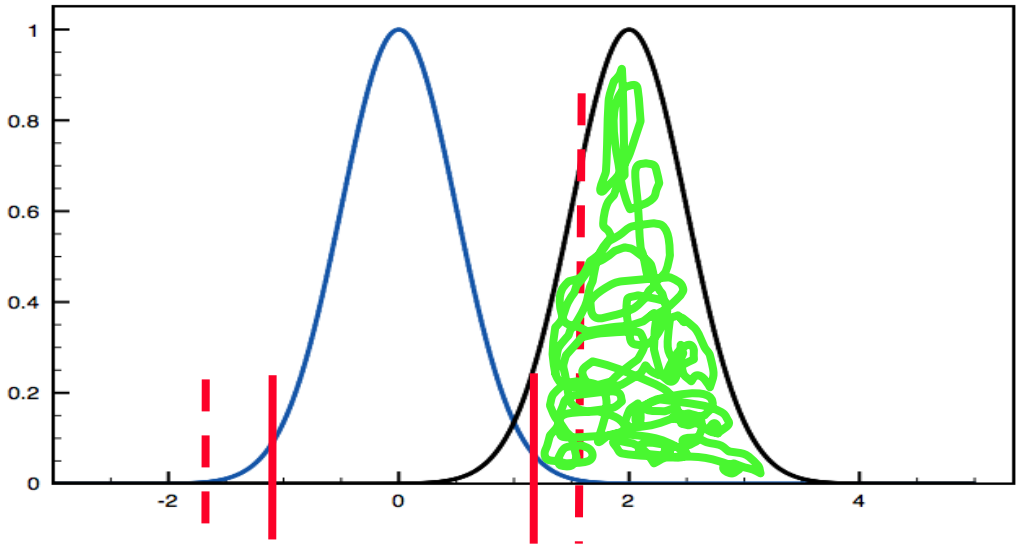
# Power

- If the alternative hypothesis is true
- Power is the probability you will reject $H_0$
- If you repeated the experiment many times, you would expect to reject $H_0$ with a proportion that reflects the power



# Power and sample size

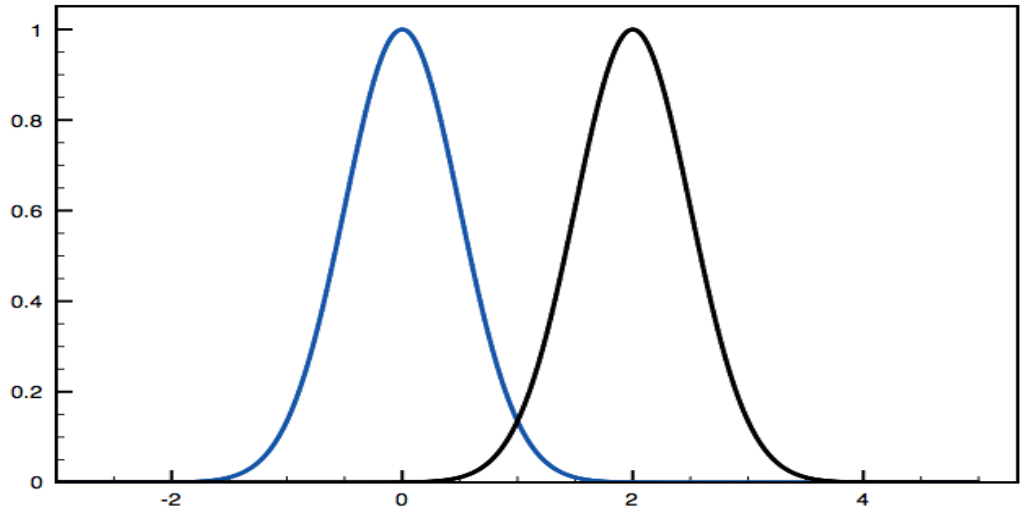
- The standard deviation of the sampling distribution is inversely related to the (square root of the) sample size
- Power increases with larger sample sizes



# Effect Size

- The difference between the null and alternative hypotheses can be characterized by a standardized effect size

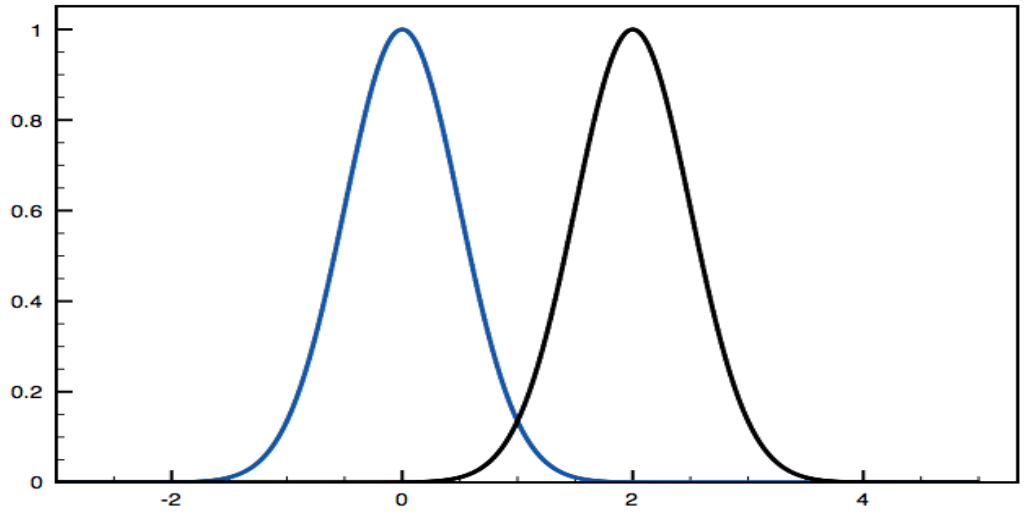
$$g = c(m)\frac{\overline{X}_1 - \overline{X}_2}{s}$$



# Effect Size

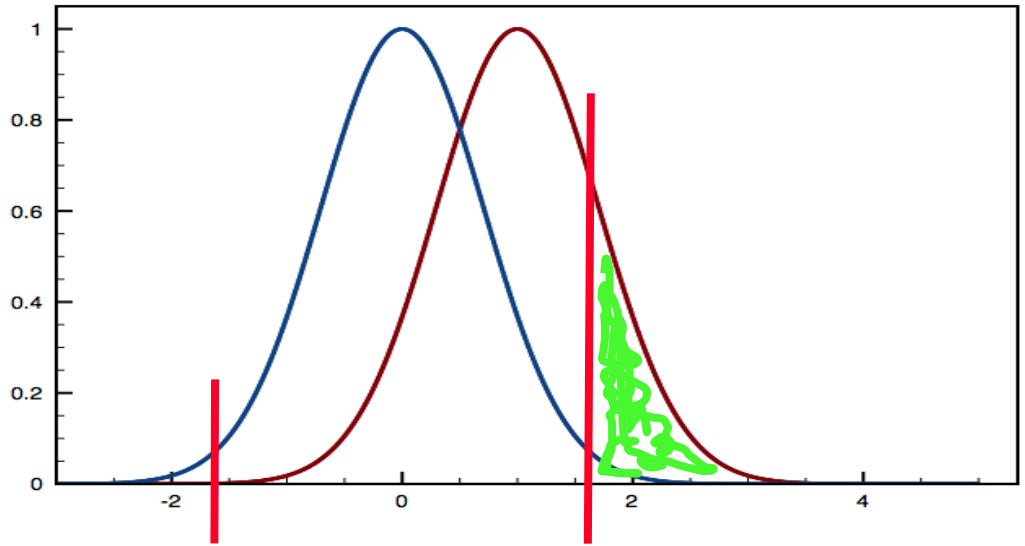
- Effect size does not vary with sample size
- although the estimate may become more accurate with larger samples

$$g = c(m)\frac{\overline{X}_1 - \overline{X}_2}{s}$$



# Effect size and power

- Experiments with smaller effect sizes have smaller power



# Effect size

- Consider the 10 findings reported by Bem (2011)
- All experiments were measured as a one-sample *t*-test (one tail, Type I error rate of 0.05)
- For each experiment, we can measure the standardized effect size (Hedges *g*)

$$g = c(m)\frac{\overline{X} - m_0}{s}$$

- Where *c(m)* is a correction for small samples sizes (≈1)
- *s* is the sample standard deviation, $\overline{X}$ is the sample mean
- $\mu_0$ is the value in the null hypothesis

# Effect size

- Use meta-analytic techniques to pool the effect sizes across all ten experiments (Hedges & Olkin, 1985)
- Pooled effect size
  - $g^*$=0.1855

$$g^* = \frac{\sum_{i=1}^{M} w_i g_i}{\sum_{i=1}^{M} w_i}$$

- $w_i$ is the inverse variance of the effect size estimate

| | Sample size | Effect size (g) |
|---|---|---|
| Exp. 1 | 100 | 0.249 |
| Exp. 2 | 150 | 0.194 |
| Exp. 3 | 97 | 0.248 |
| Exp. 4 | 99 | 0.202 |
| Exp. 5 | 100 | 0.221 |
| Exp. 6 Negative | 150 | 0.146 |
| Exp. 6 Erotic | 150 | 0.144 |
| Exp. 7 | 200 | 0.092 |
| Exp. 8 | 100 | 0.191 |
| Exp. 9 | 50 | 0.412 |

# Power

- Use the pooled effect size to compute the power of each experiment (probability this experiment would reject the null hypothesis)
- Pooled effect size
  - $g^*$=0.1855

| | Sample size | Effect size (g) | Power |
|---|---|---|---|
| Exp. 1 | 100 | 0.249 | 0.578 |
| Exp. 2 | 150 | 0.194 | 0.731 |
| Exp. 3 | 97 | 0.248 | 0.567 |
| Exp. 4 | 99 | 0.202 | 0.575 |
| Exp. 5 | 100 | 0.221 | 0.578 |
| Exp. 6 Negative | 150 | 0.146 | 0.731 |
| Exp. 6 Erotic | 150 | 0.144 | 0.731 |
| Exp. 7 | 200 | 0.092 | 0.834 |
| Exp. 8 | 100 | 0.191 | 0.578 |
| Exp. 9 | 50 | 0.412 | 0.363 |

# Power

- The sum of the power values ($E$=6.27) is the *expected* number of times experiments like these would reject the null hypothesis
  (Ioannidis & Trikalinos, 2007)
- But Bem (2011) rejected the null $O$=9 out of 10 times!

| | Sample size | Effect size (g) | Power |
|---|---|---|---|
| Exp. 1 | 100 | 0.249 | 0.578 |
| Exp. 2 | 150 | 0.194 | 0.731 |
| Exp. 3 | 97 | 0.248 | 0.567 |
| Exp. 4 | 99 | 0.202 | 0.575 |
| Exp. 5 | 100 | 0.221 | 0.578 |
| Exp. 6 Negative | 150 | 0.146 | 0.731 |
| Exp. 6 Erotic | 150 | 0.144 | 0.731 |
| Exp. 7 | 200 | 0.092 | 0.834 |
| Exp. 8 | 100 | 0.191 | 0.578 |
| Exp. 9 | 50 | 0.412 | 0.363 |

# Bias Test

- Use an exact test to consider the probability that any *O*=9 out of the 10 experiments would reject $H_0$
- There are 11 such combinations of the experiments
- Their summed probability is only 0.058
- A criterion threshold for a bias test is usually 0.1
  (Begg & Mazumdar, 1994; Ioannidis & Trikalinos, 2007; Stern & Egger, 2001)

| | Sample size | Effect size (g) | Power |
|---|---|---|---|
| Exp. 1 | 100 | 0.249 | 0.578 |
| Exp. 2 | 150 | 0.194 | 0.731 |
| Exp. 3 | 97 | 0.248 | 0.567 |
| Exp. 4 | 99 | 0.202 | 0.575 |
| Exp. 5 | 100 | 0.221 | 0.578 |
| Exp. 6 Negative | 150 | 0.146 | 0.731 |
| Exp. 6 Erotic | 150 | 0.144 | 0.731 |
| Exp. 7 | 200 | 0.092 | 0.834 |
| Exp. 8 | 100 | 0.191 | 0.578 |
| Exp. 9 | 50 | 0.412 | 0.363 |

# Interpretation

- The number of times Bem (2011) rejected the $H_0$ is inconsistent with the size of the reported effect and the properties of the experiments
  1. Perhaps there were additional experiments that failed to reject $H_0$ but were not reported
  2. Perhaps the experiments were run incorrectly in a way that rejected the $H_0$ too frequently
  3. Perhaps the experiments were run incorrectly in a way that underestimated the true magnitude of the effect size
- The findings in Bem (2011) seem *too good to be true*
  - Non-scientific set of findings
  - Anecdotal
- Note, the effect may be true (or not), but the studies in Bem (2011) give no guidance

# Bystander Effect

Fischer *et al.* (2011) described a meta-analysis of studies of the bystander effect

Broke down studies according to emergency or non-emergency situations

# Bystander Effect

- No suspicion of publication bias for *non-emergency* situations
  - Effect "B" from the earlier slides
- Clear indication of publication bias for *emergency situations*
  - Even though fewer than half of the experiments reject $H_0$

| | Emergency situation | Non-emergency situation |
|---|---|---|
| Number of studies | 65 | 19 |
| Pooled effect size | -0.30 | -0.47 |
| Observed number of rejections of $H_0$ consistent with bystander effect (O) | 24 | 10 |
| Expected number of rejections of $H_0$ consistent with bystander effect (E) | 10.02 | 10.77 |
| $\chi^2(1)$ | 23.05 | 0.128 |
| p | <.0001 | 0.721 |

# Simulated Replications

- Two-sample *t* test
- Control group: draw $n_1$ samples from a normal distribution N(0,1)
- Experimental group: drawn $n_2=n_1$ samples from a normal distribution N(0.3,1)
  - The true effect size is 0.3
- Repeat for 20 experiments
  - With random samples sizes $n_2=n_1$ drawn uniformly from [15, 50]

# Simulated Replications

- Compute the pooled effect size
  - $g^*=0.303$
  - Very close to true 0.3

| $n_1=n_2$ | t | Effect size |
|---|---|---|
| 29 | 0.888 | 0.230 |
| 25 | 1.380 | 0.384 |
| 26 | 1.240 | 0.339 |
| 15 | 0.887 | 0.315 |
| 42 | 0.716 | 0.155 |
| 37 | 1.960 | 0.451 |
| 49 | -0.447 | -0.090 |
| 17 | 1.853 | 0.621 |
| 36 | 2.036 | 0.475 |
| 22 | 1.775 | 0.526 |
| 39 | 1.263 | 0.283 |
| 19 | 3.048 | 0.968 |
| 18 | 2.065 | 0.673 |
| 26 | -1.553 | -0.424 |
| 38 | -0.177 | -0.040 |
| 42 | 2.803 | 0.606 |
| 21 | 1.923 | 0.582 |
| 40 | 2.415 | 0.535 |
| 22 | 1.786 | 0.529 |
| 35 | -0.421 | -0.100 |

# Simulated Replications

- Compute the pooled effect size
  - $g^*$=0.303
  - Very close to true 0.3
- Use effect size to compute power
- Sum of power is expected number of times to reject
  - *E*(true)=4.140
  - *E*(pooled)=4.214
- Observed rejections
  - *O*=5

| $n_1=n_2$ | t | Effect size | Power from true ES | Power from pooled ES |
|---|---|---|---|---|
| 29 | 0.888 | 0.230 | 0.202 | 0.206 |
| 25 | 1.380 | 0.384 | 0.180 | 0.183 |
| 26 | 1.240 | 0.339 | 0.186 | 0.189 |
| 15 | 0.887 | 0.315 | 0.125 | 0.126 |
| 42 | 0.716 | 0.155 | 0.274 | 0.279 |
| 37 | 1.960 | 0.451 | 0.247 | 0.251 |
| 49 | -0.447 | -0.090 | 0.312 | 0.318 |
| 17 | 1.853 | 0.621 | 0.136 | 0.138 |
| 36 | 2.036 | 0.475 | 0.241 | 0.245 |
| 22 | 1.775 | 0.526 | 0.163 | 0.166 |
| 39 | 1.263 | 0.283 | 0.258 | 0.262 |
| 19 | 3.048 | 0.968 | 0.147 | 0.149 |
| 18 | 2.065 | 0.673 | 0.141 | 0.143 |
| 26 | -1.553 | -0.424 | 0.186 | 0.189 |
| 38 | -0.177 | -0.040 | 0.252 | 0.257 |
| 42 | 2.803 | 0.606 | 0.274 | 0.279 |
| 21 | 1.923 | 0.582 | 0.158 | 0.160 |
| 40 | 2.415 | 0.535 | 0.263 | 0.268 |
| 22 | 1.786 | 0.529 | 0.163 | 0.166 |
| 35 | -0.421 | -0.100 | 0.236 | 0.240 |
| | | | Σ = 4.14 | 4.214 |

# Simulated Replications

- Probability of observing $O \geq 5$ rejections for 20 experiments like these is
  - 0.407 for true ES
  - 0.417 for pooled ES
- No indication of publication bias when all the experiments are fully reported

| $n_1=n_2$ | t | Effect size | Power from true ES | Power from pooled ES |
|---|---|---|---|---|
| 29 | 0.888 | 0.230 | 0.202 | 0.206 |
| 25 | 1.380 | 0.384 | 0.180 | 0.183 |
| 26 | 1.240 | 0.339 | 0.186 | 0.189 |
| 15 | 0.887 | 0.315 | 0.125 | 0.126 |
| 42 | 0.716 | 0.155 | 0.274 | 0.279 |
| 37 | 1.960 | 0.451 | 0.247 | 0.251 |
| 49 | -0.447 | -0.090 | 0.312 | 0.318 |
| 17 | 1.853 | 0.621 | 0.136 | 0.138 |
| 36 | 2.036 | 0.475 | 0.241 | 0.245 |
| 22 | 1.775 | 0.526 | 0.163 | 0.166 |
| 39 | 1.263 | 0.283 | 0.258 | 0.262 |
| 19 | 3.048 | 0.968 | 0.147 | 0.149 |
| 18 | 2.065 | 0.673 | 0.141 | 0.143 |
| 26 | -1.553 | -0.424 | 0.186 | 0.189 |
| 38 | -0.177 | -0.040 | 0.252 | 0.257 |
| 42 | 2.803 | 0.606 | 0.274 | 0.279 |
| 21 | 1.923 | 0.582 | 0.158 | 0.160 |
| 40 | 2.415 | 0.535 | 0.263 | 0.268 |
| 22 | 1.786 | 0.529 | 0.163 | 0.166 |
| 35 | -0.421 | -0.100 | 0.236 | 0.240 |
| | | | Σ = 4.14 | 4.214 |

# Simulated File Drawer

- Suppose a researcher only published the experiments that rejected the null hypothesis
- The pooled effect size is now
  - $g^*=0.607$
  - Double the true effect!
- Also increases the estimated power of the reported experiments

| $n_1=n_2$ | t | Effect size | Power from true ES | Power from pooled ES |
|---|---|---|---|---|
| 29 | 0.888 | 0.230 | 0.202 | 0.206 |
| 25 | 1.380 | 0.384 | 0.180 | 0.183 |
| 26 | 1.240 | 0.339 | 0.186 | 0.189 |
| 15 | 0.887 | 0.315 | 0.125 | 0.126 |
| 42 | 0.716 | 0.155 | 0.274 | 0.279 |
| 37 | 1.960 | 0.451 | 0.247 | 0.251 |
| 49 | -0.447 | -0.090 | 0.312 | 0.318 |
| 17 | 1.853 | 0.621 | 0.136 | 0.138 |
| 36 | 2.036 | 0.475 | 0.241 | 0.245 |
| 22 | 1.775 | 0.526 | 0.163 | 0.166 |
| 39 | 1.263 | 0.283 | 0.258 | 0.262 |
| 19 | 3.048 | 0.968 | 0.147 | 0.149 |
| 18 | 2.065 | 0.673 | 0.141 | 0.143 |
| 26 | -1.553 | -0.424 | 0.186 | 0.189 |
| 38 | -0.177 | -0.040 | 0.252 | 0.257 |
| 42 | 2.803 | 0.606 | 0.274 | 0.279 |
| 21 | 1.923 | 0.582 | 0.158 | 0.160 |
| 40 | 2.415 | 0.535 | 0.263 | 0.268 |
| 22 | 1.786 | 0.529 | 0.163 | 0.166 |
| 35 | -0.421 | -0.100 | 0.236 | 0.240 |
| | | | Σ = 4.14 | 4.214 |

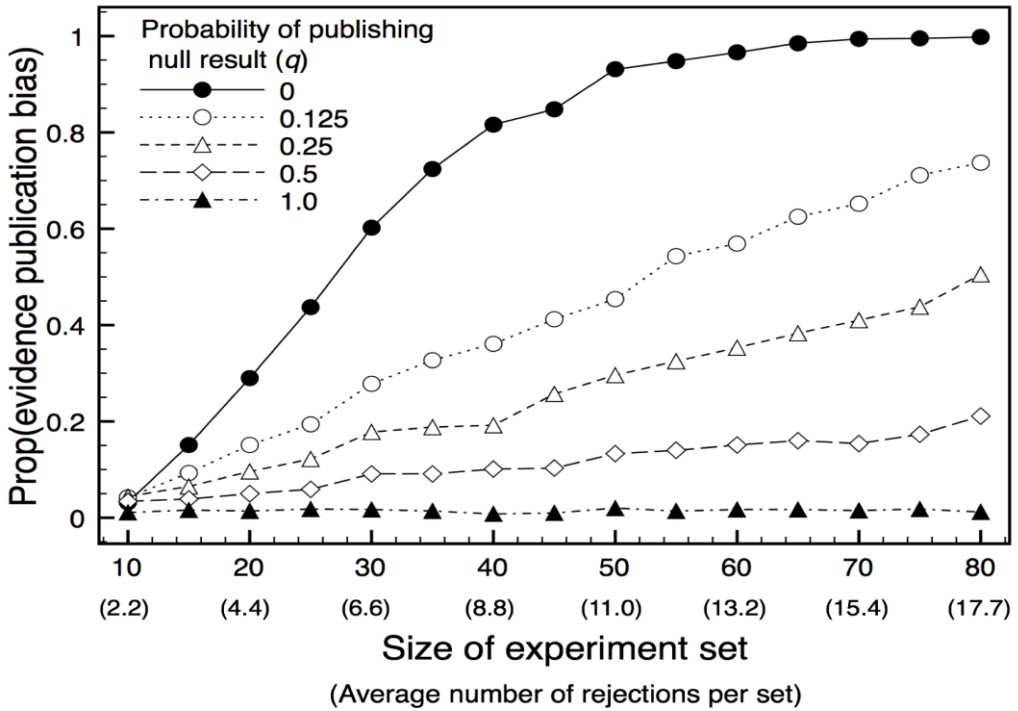
# Simulated File Drawer

- The sum of power values is again the expected number of times the null hypothesis should be rejected
  - E(biased)=3.135
  - Compare to O=5
- The probability of 5 experiments like these all rejecting the null is the product of the power terms
  - 0.081 (<0.1)
  - Indicates publication bias

| $n_1=n_2$ | t | Effect size | Power from true ES | Power from pooled ES | Power from biased ES |
|---|---|---|---|---|---|
| 29 | 0.888 | 0.230 | 0.202 | 0.206 | |
| 25 | 1.380 | 0.384 | 0.180 | 0.183 | |
| 26 | 1.240 | 0.339 | 0.186 | 0.189 | |
| 15 | 0.887 | 0.315 | 0.125 | 0.126 | |
| 42 | 0.716 | 0.155 | 0.274 | 0.279 | |
| 37 | 1.960 | 0.451 | 0.247 | 0.251 | |
| 49 | -0.447 | -0.090 | 0.312 | 0.318 | |
| 17 | 1.853 | 0.621 | 0.136 | 0.138 | |
| 36 | 2.036 | 0.475 | 0.241 | 0.245 | 0.718 |
| 22 | 1.775 | 0.526 | 0.163 | 0.166 | |
| 39 | 1.263 | 0.283 | 0.258 | 0.262 | |
| 19 | 3.048 | 0.968 | 0.147 | 0.149 | 0.444 |
| 18 | 2.065 | 0.673 | 0.141 | 0.143 | 0.424 |
| 26 | -1.553 | -0.424 | 0.186 | 0.189 | |
| 38 | -0.177 | -0.040 | 0.252 | 0.257 | |
| 42 | 2.803 | 0.606 | 0.274 | 0.279 | 0.784 |
| 21 | 1.923 | 0.582 | 0.158 | 0.160 | |
| 40 | 2.415 | 0.535 | 0.263 | 0.268 | 0.764 |
| 22 | 1.786 | 0.529 | 0.163 | 0.166 | |
| 35 | -0.421 | -0.100 | 0.236 | 0.240 | |
| | | | Σ = 4.14 | 4.214 | 3.135 |

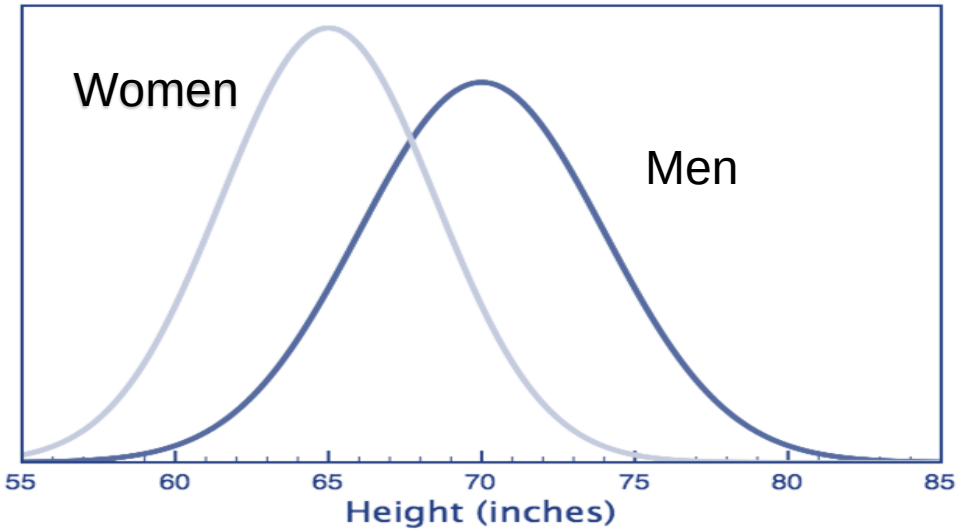
# Simulated File Drawer

- The test for publication bias works properly
- But it is conservative
- If the test indicates bias, we can be fairly confident it is correct



# Statistical Errors

- Even if an effect is truly zero, a random sample will sometimes produce a significant effect (false alarm: α)
- Even if an effect is non-zero, a random sample will not always produce a statistically significant effect (miss: β=1-power)
- A scientist who does not sometimes make a mistake with statistics is *doing it wrong*
- There can be **excess success**



# Simulated Optional Stopping

- There are other types of biases
- Set true effect size to 0
- Optional stopping:
  - Take sample of $n_1=n_2=15$
  - Run hypothesis test
  - If reject null or $n_1=n_2=100$, stop and report
  - Otherwise, add one more sample to each group and repeat
- Just by random sampling, *O*=4 experiments reject the null hypothesis
  - Type I error rate of 0.2, even though used $\alpha=0.05$

| $n_1=n_2$ | t | Effect size |
|---|---|---|
| **19** | **2.393** | **0.760** |
| 100 | 0.774 | 0.109 |
| 100 | 1.008 | 0.142 |
| **63** | **2.088** | **0.370** |
| 100 | 0.587 | 0.083 |
| 100 | -1.381 | -0.195 |
| 100 | -0.481 | -0.068 |
| 100 | 0.359 | 0.051 |
| 100 | -1.777 | -0.250 |
| 100 | -0.563 | -0.079 |
| 100 | 1.013 | 0.143 |
| 100 | -0.012 | -0.002 |
| **46** | **2.084** | **0.431** |
| 100 | 0.973 | 0.137 |
| 100 | -0.954 | -0.134 |
| 100 | -0.136 | -0.019 |
| **78** | **2.052** | **0.327** |
| 100 | -0.289 | -0.041 |
| 100 | 1.579 | 0.222 |
| 100 | 0.194 | 0.027 |

# Simulated Optional Stopping

- Pooled effect size across all experiments is $g^*=0.052$
  - Sum of power values is $E=1.28$
  - Probability of $O \geq 4$ is 0.036

| $n_1=n_2$ | t | Effect size | Power from pooled ES |
|---|---|---|---|
| **19** | **2.393** | **0.760** | 0.053 |
| 100 | 0.774 | 0.109 | 0.066 |
| 100 | 1.008 | 0.142 | 0.066 |
| **63** | **2.088** | **0.370** | 0.060 |
| 100 | 0.587 | 0.083 | 0.066 |
| 100 | -1.381 | -0.195 | 0.066 |
| 100 | -0.481 | -0.068 | 0.066 |
| 100 | 0.359 | 0.051 | 0.066 |
| 100 | -1.777 | -0.250 | 0.066 |
| 100 | -0.563 | -0.079 | 0.066 |
| 100 | 1.013 | 0.143 | 0.066 |
| 100 | -0.012 | -0.002 | 0.066 |
| **46** | **2.084** | **0.431** | 0.057 |
| 100 | 0.973 | 0.137 | 0.066 |
| 100 | -0.954 | -0.134 | 0.066 |
| 100 | -0.136 | -0.019 | 0.066 |
| **78** | **2.052** | **0.327** | 0.062 |
| 100 | -0.289 | -0.041 | 0.066 |
| 100 | 1.579 | 0.222 | 0.066 |
| 100 | 0.194 | 0.027 | 0.066 |

$\Sigma = 1.28$

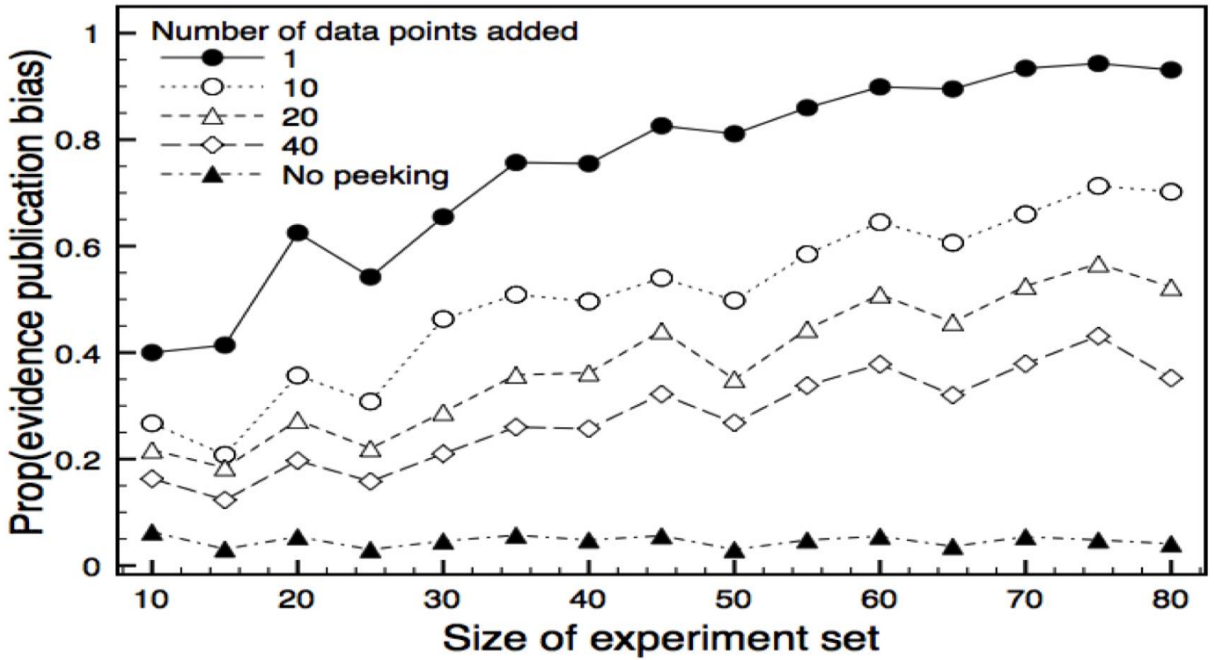
# Simulated Optional Stopping

- Pooled effect size across all experiments is $g^*$=0.052
  - Sum of power values is *E*=1.28
  - Probability of *O*≥4 is 0.036
- If add a file-drawer bias
  - $g^*$=0.402
  - *E*=2.02
  - *P*=0.047

| $n_1=n_2$ | t | Effect size | Power from pooled ES | Power from file drawer ES |
|---|---|---|---|---|
| **19** | **2.393** | **0.760** | **0.053** | **0.227** |
| 100 | 0.774 | 0.109 | 0.066 | |
| 100 | 1.008 | 0.142 | 0.066 | |
| **63** | **2.088** | **0.370** | **0.060** | **0.611** |
| 100 | 0.587 | 0.083 | 0.066 | |
| 100 | -1.381 | -0.195 | 0.066 | |
| 100 | -0.481 | -0.068 | 0.066 | |
| 100 | 0.359 | 0.051 | 0.066 | |
| 100 | -1.777 | -0.250 | 0.066 | |
| 100 | -0.563 | -0.079 | 0.066 | |
| 100 | 1.013 | 0.143 | 0.066 | |
| 100 | -0.012 | -0.002 | 0.066 | |
| **46** | **2.084** | **0.431** | **0.057** | **0.480** |
| 100 | 0.973 | 0.137 | 0.066 | |
| 100 | -0.954 | -0.134 | 0.066 | |
| 100 | -0.136 | -0.019 | 0.066 | |
| **78** | **2.052** | **0.327** | **0.062** | **0.704** |
| 100 | -0.289 | -0.041 | 0.066 | |
| 100 | 1.579 | 0.222 | 0.066 | |
| 100 | 0.194 | 0.027 | 0.066 | |
| | | | Σ = 1.28 | 2.02 |

# Simulated  Optional Stopping bias

- The test for publication bias works properly
- But it is conservative
- When the test indicates bias, it is almost always correct



# Data And Theory

- Elliot, Niesta Kayser, Greitemeyer, Lichtenfeld, Gramzow, Maier & Liu (2010). Red, rank, and romance in women viewing men. *Journal of Experimental Psychology: General*
- Picked up by the popular press

# Data And Theory

- 7 successful experiments, three theoretical conclusions

  1) Women perceive men to be more attractive when seen on a red background and in red clothing

  2) Women perceive men to be more sexually desirable when seen on a red background and in red clothing

  3) Changes in perceived status are responsible for these effects

# Analysis: Attractiveness

- Pooled effect size is
  - $g^*$=0.785
- Every reported experiment rejected the null
- Given the power values, the expected number of rejections is *E*=2.86
- The estimated probability of five experiments like these to all reject the null is 0.054

| Description | N1 | N2 | Effect size | Power from pooled ES |
|---|---|---|---|---|
| Exp. 1 | 10 | 11 | 0.914 | .400 |
| Exp. 2 | 20 | 12 | 1.089 | .562 |
| Exp. 3 | 16 | 17 | 0.829 | .589 |
| Exp. 4 | 27 | 28 | 0.54 | .816 |
| Exp. 7 | 12 | 15 | 0.824 | .496 |

# Analysis: Desirability

- Pooled effect size is
  - $g^*$=0.744
- Every reported experiment rejected the null
- The estimated probability of three experiments like these to all reject the null is 0.191

| Description | N1 | N2 | Effect size | Power from pooled ES |
|---|---|---|---|---|
| Exp. 3 | 16 | 17 | 0.826 | .544 |
| Exp. 4 | 27 | 28 | 0.598 | .773 |
| Exp. 7 | 12 | 15 | 0.952 | .455 |

# Analysis: Status

- Pooled effect size is
  - $g^*$=0.894
- Every reported experiment rejected the null
- The estimated probability of three experiments like these to all reject the null is 0.179

| Description | N1 | N2 | Effect size | Power from pooled ES |
|---|---|---|---|---|
| Exp. 5a (present) | 10 | 10 | .929 | .395 |
| Exp. 5a (potential) | 10 | 10 | 1.259 | |
| Exp. 6 | 19 | 18 | 0.718 | .752 |
| Exp. 7 | 12 | 15 | 0.860 | .602 |

# Future Studies

- The probabilities for desirability and status do not fall below the 0.1 threshold
- One more successful experimental result for these measures is likely to drop the power probability below the criterion
- These results will be most believable if a replication *fails* to show a statistically significant result
  - But just barely fails
  - A convincing fail will have a small effect size, which will pull down the estimated power of the other studies

# Theories From Data

- Elliot et al. (2010) proposed a theory
  - Red influences perceived status, which then influences perceived attractiveness and desirability
- Such a claim requires (at least) that all three results be valid
- Several experiments measured these variables with a single set of subjects
  - The data on these measures are correlated
  - Total power is not just the product of probabilities
  - Can recalculate power with provided correlations among variables

# Analysis: Correlated Data

- Every reported test rejected the null
- The estimated probability of 12 hypothesis tests in seven experiments like these to all reject the null is 0.005

| Description | Power from pooled ES |
|---|---|
| Exp. 1, Attractiveness, desirability | .400 |
| Exp. 2, Attractiveness | .562 |
| Exp. 3, Attractiveness, desirability | .438 |
| Exp. 4, Attractiveness, desirability | .702 |
| Exp. 5a, Status | .395 |
| Exp. 6 Status | .752 |
| Exp. 7 Attractiveness, desirability, status | .237 |

# Theories From Data

- Elliot et al. (2010) proposed a theory
  - Red influences perceived status, which then influences perceived attractiveness and desirability
- This theory also generated five predicted null findings
  - E.g., Men do not show the effect of perceived attractiveness when rating other men
- If the null is true for these cases, the probability of all five tests not rejecting the null is
  - $(1 - 0.05)^5 = 0.77$
- The theory never made a mistake in predicting the outcome of a hypothesis test
  - The estimated probability of such an outcome is
  - 0.005 x 0.77 = 0.0038

# Response From Elliot & Maier (2013)

- Lots of other labs have verified the red-attractiveness effect
  - If these other studies form part of the evidence for their theory, they only strengthen the claim of bias (which now includes other labs)
- Conducted a replication study of Experiment 3
  - $N_1$=75 women judged attractiveness of men's photos with red
  - $N_2$=69 women judged attractiveness of men's photos with gray
  - Results: t= 1.51, p=.13, effect size = 0.25
- They conclude that the effect is real, but smaller than they originally estimated
  - Implies that they do not believe in hypothesis testing.

# Analysis: Attractiveness 2

- Pooled effect size is
  - $g^*$=~~0.785~~ 0.532
- Given the power values, the expected number of rejections is *E*=~~2.86~~ 2.47
- The estimated probability of five out of ~~five~~ six experiments like these to reject the null is ~~0.054~~ 0.030

| Description | N1 | N2 | Effect size | Power from pooled ES |
|---|---|---|---|---|
| Exp. 1 | 10 | 11 | 0.914 | ~~.400~~ .212 |
| Exp. 2 | 20 | 12 | 1.089 | ~~.562~~ .297 |
| Exp. 3 | 16 | 17 | 0.829 | ~~.589~~ .316 |
| Exp. 4 | 27 | 28 | 0.54 | ~~.816~~ .491 |
| Exp. 7 | 12 | 15 | 0.824 | ~~.496~~ .262 |
| Replication | 75 | 69 | 0.251 | .887 |

# Analysis: Attractiveness 2’

- One could argue that the best estimate of the effect is from the replication experiment
  - $g^*$=~~0.785~~ ~~0.532~~ 0.251
- Given the power values, the expected number of rejections is *E*=~~2.86~~ ~~2.47~~ 0.860
- The estimated probability of five out of ~~five~~ six experiments like these to reject the null is ~~0.054~~ ~~0.030~~ 0.0002
- The estimated probability of the original 5 experiments all being successful is 0.000013

| Description | N1 | N2 | Effect size | Power from pooled ES |
|---|---|---|---|---|
| Exp. 1 | 10 | 11 | 0.914 | ~~.400~~ ~~.212~~ .085 |
| Exp. 2 | 20 | 12 | 1.089 | ~~.562~~ ~~.297~~ .103 |
| Exp. 3 | 16 | 17 | 0.829 | ~~.589~~ ~~.316~~ .107 |
| Exp. 4 | 27 | 28 | 0.54 | ~~.816~~ ~~.491~~ .149 |
| Exp. 7 | 12 | 15 | 0.824 | ~~.496~~ ~~.262~~ .095 |
| Replication | 75 | 69 | 0.251 | ~~.887~~ .320 |

# Analysis: Attractiveness 3

- A recent meta-analysis (*n*=3,381) (Lehmann, Elliot, Calin-Jageman, 2018)
  - Finds small effect size (d=0.13)
  - Evidence of publication bias
- Two conclusions sections:
  - **First and Third Authors:** "The simplest conclusion from our results is that the true effect of incidental red on attraction is very small, potentially nonexistent."
  - **Second Author:** "Two primary weaknesses are that nearly all existing studies are underpowered and fail to attend to important color science procedures, especially regarding color production (e.g., spectral assessment, matching color attributes) and presentation (e.g., ambient illumination, background contrast; Elliot, 2015; Fairchild, 2015). Indeed, not a single published study that contributed to our main meta-analysis would be considered exemplary based on these two criteria alone."

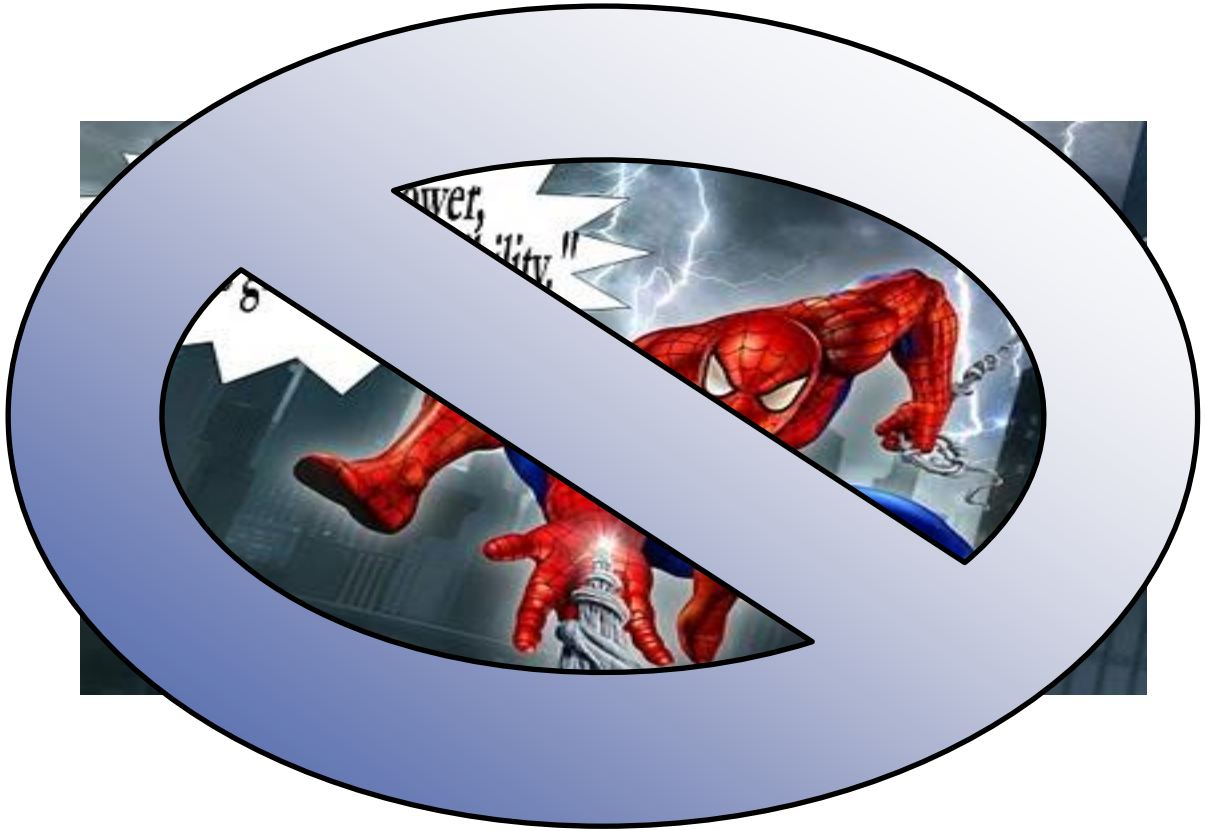
# Power And Replication

- Studies that depend on hypothesis testing can only detect a given effect with a certain probability
  - Due to random sampling
- Even if the effect is true, you should sometimes fail to reject $H_0$
- *The frequency of rejecting $H_0$ must reflect the underlying power of the experiments*
  - When the observed number of rejections is radically different from what is to be expected, something is wrong (publication bias, optional stopping, something else)

# Good News

- Many people get very concerned when their experimental finding is not replicated by someone else
- Lots of accusations about incompetence and suppositions about who is wrong
- But "failure" to replicate is *expected* when decisions are made with hypothesis testing
  - At a rate dependent on the experimental power
- Statisticians have an *obligation* to be **wrong** the specified proportion of the time

# Conclusions



# Conclusions

- For a scientist, ***low*** power comes with great responsibility
  - A scientist must resist the temptation to make the data show what they desire
  - A scientist must not keep gathering data until finding what he desires
  - A scientist must not try different analyses until finding one that shows what she wants.
  - A scientist must resist drawing firm conclusions from noisy data
- Being a responsible scientist is easy when the signal is clear and noise is small (high power)
- Statistics is easy with large power.

# Take home messages

**Take Home Messages**

1. If many similar experiments with low effect and sample size all lead to significant results: the data seem too good to be true.
2. Experiments should lead to significant results proportional to their power.
3. Publication bias and optional stopping can lead to strongly inflated Type I error rates.

# END Class 10