What about Design 3?

		Sit		Stand		Sit	# people	# measures	#EU
3)	Jill	X T1 X T2	Bob	X T1 X T2	Amy	X T1 X T2	40	80	40

$$\sigma_r(\hat{\delta}) = \sqrt{\sigma_r^2(\hat{\mu}_B) + \sigma_r^2(\hat{\mu}_A)}$$ **Indirect** Standard Error

$$\sigma_r(\hat{\mu}_i) = \sqrt{\frac{\text{Variance of population + Variance of measurements}}{\text{Sample size}}}$$ **Direct** Standard Error

population = TRUE **Standing pulse** μ_{Aj}

population variance = $\sigma^2_{\mu_i}$ (same as Design 1)

measurements = Estimates of Standing pulse $\hat{\mu}_{Aj} = \dfrac{y_{Aj_1} + y_{Aj_2}}{2}$

measurement variance is 1/2 that of Design 1: $\dfrac{\sigma_m^2}{2}$

samples / treatment level is 1/2 that of design 1

Final Standard Error:

$$\sigma_r(\hat{\delta}) = \sqrt{\frac{2\sigma^2_{\mu_i} + 2/2\sigma^2_m}{20}}$$

Efficiency of Experimental Designs

Design 2

$$\sigma_r(\hat{\delta}) = \sqrt{\frac{\sigma_\delta^2 + 2\sigma_m^2}{40}}$$

Smaller population variance

Design 1

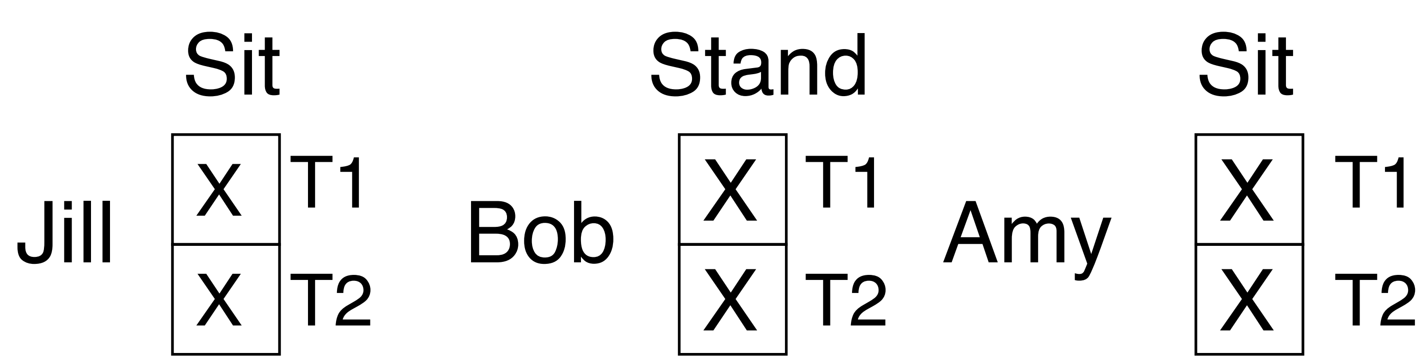
$$\sigma_r(\hat{\delta}) = \sqrt{\frac{2\sigma_{\mu_i}^2 + 2\sigma_m^2}{40}}$$

Smaller measurement error

Design 3

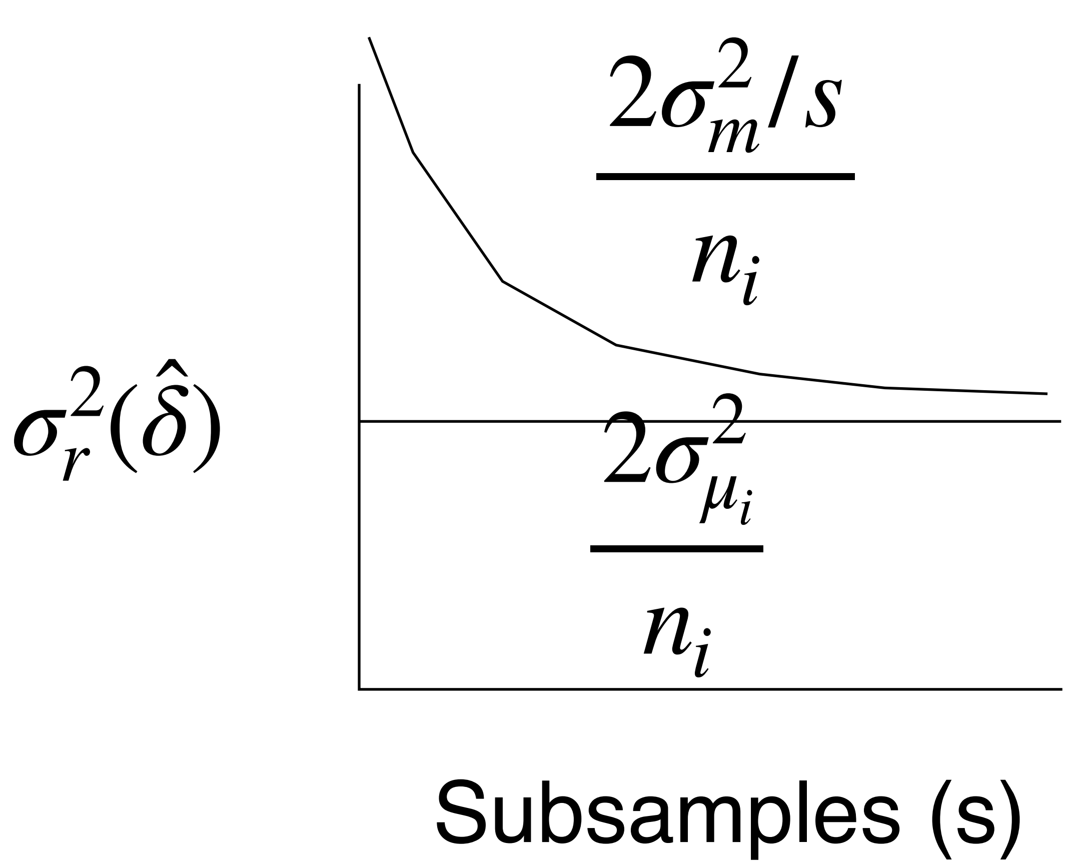
$$\sigma_r(\hat{\delta}) = \sqrt{\frac{2\sigma_{\mu_i}^2 + 2/2\sigma_m^2}{20}}$$

Smaller sample size (#EU / level)



Repeated measures of the same sample are **subsamples**

Not declaring EU causes **pseudoreplication**



More subsamples = less measurement error / EU

More subsamples = fewer EU

Optimal number of subsamples

Tradeoff between EU and subsamples based on cost

Optimal number of subsamples $\sqrt{\frac{k}{c}}$

$$c = \frac{\text{cost/(subsample + measurement)}}{\text{cost/(EU)}}$$

$$k = \frac{\sigma_m^2}{\sigma_{\mu_i}^2} \quad \left(\frac{\dfrac{\hat{\mu}_{ij1} - \hat{\mu}_{ij2}}{2}}{\dfrac{\mu_{i1} - \mu_{i2}}{2}} \right)^2$$

$\dfrac{\hat{\mu}_{ij1} - \hat{\mu}_{ij2}}{2}$: half the difference among replicate measures of the same individual

$\dfrac{\mu_{i1} - \mu_{i2}}{2}$: half the difference among different individuals

k	c	Optimal number of subsamples
.5	1	.7 $\longrightarrow$ 1
1	1	1 $\longrightarrow$ 1
2	1	1.4 $\longrightarrow$ 1
4	1	2 $\longrightarrow$ 2
.5	1/10	2.2 $\longrightarrow$ 2
1	1/10	3.1 $\longrightarrow$ 3
2	1/10	4.5 $\longrightarrow$ 4

Estimated Standard Errors for each design

Design 2 "Direct"

$$\hat{\delta}_{B-A} = \frac{1}{n}\sum \hat{\delta}_j$$

Direct estimate = average of *n* observations

TRUE Standard Error:

$$\sigma_r(\hat{\delta}) = \sqrt{\frac{\text{Variance of population + Variance of measurements}}{\text{Sample size}}}$$

Estimated Standard Error:

Sample Variance of $\hat{\delta}_i \approx$ Variance of TRUE values + Variance of erro

$$s^2_{\hat{\delta}} = \frac{\sum(\hat{\delta}_j - \hat{\delta})^2}{n-1}$$

SED = $\sqrt{\frac{s^2_{\hat{\delta}}}{n}}$ This is an estimate of $\sigma_r(\hat{\delta})$

Degrees of Freedom

(n-1) from denominator of $s^2_{\hat{\delta}}$

Estimated Standard Errors for each design

Design 1 "Indirect"

$$\hat{\delta}_{B-A} = \hat{\mu}_B - \hat{\mu}_A$$ Indirect estimate of δ

$$\hat{\mu}_A = \frac{1}{n_A} \sum \hat{\mu}_{Aj}$$ direct estimates of μ_B and μ_A

TRUE Standard Error:

$$\sigma_r(\hat{\mu}_i) = \sqrt{\frac{\text{Variance of population + Variance of measurements}}{\text{Sample size}}}$$

Estimated Standard Error:

Sample Variance of $\hat{\mu}_{ij} \approx$ Variance of TRUE values + Variance of error

$$s^2_{\hat{\mu}_i} = \frac{\sum (\hat{\mu}_{ij} - \hat{\mu}_i)^2}{n_i - 1}$$ **Observed** variance of estimates around their mean

SEM = $\sqrt{\frac{s^2_{\hat{\mu}_i}}{n_i}}$

TRUE Standard Error:

$$\sigma_r(\hat{\delta}) = \sqrt{\sigma_r^2(\hat{\mu}_B) + \sigma_r^2(\hat{\mu}_A)}$$

Estimated Standard Error:

SED = $\sqrt{\frac{s^2_{\hat{\mu}_B}}{n_B} + \frac{s^2_{\hat{\mu}_A}}{n_A}}$

Problem:

Degrees of Freedom $(n_B - 1)$ from $s^2_{\hat{\mu}_B}$ or $(n_A - 1)$ from $s^2_{\hat{\mu}_A}$?

Can't use 2 degrees of freedom for the t-distribution

Best Df is closer to the **smaller** of $(n_B - 1)$ and $(n_A - 1)$

Solution: **Pooled** $s^2_{\hat{\mu}}$

Sample Variance of $\hat{\mu}_{ij} \approx$ Variance of TRUE values + Variance of errors

if Variances of pulses are similar for both treatments

And measurement errors are similar

We can **pool** all deviations together into a **pooled** $s^2_{\hat{\mu}}$

$$s^2_{\hat{\mu}} = \frac{\sum(\hat{\mu}_{Bj} - \hat{\mu}_B)^2 + \sum(\hat{\mu}_{Aj} - \hat{\mu}_A)^2}{(n_B - 1) + (n_A - 1)}$$

All deviations² from the sample means

---

independent deviations per treatment

Estimated Standard Error:

SED = $\sqrt{\frac{s^2_{\hat{\mu}}}{n_B} + \frac{s^2_{\hat{\mu}}}{n_A}}$

If variances are equal, this is a **better** (more accurate) estimate of $\sigma_r(\hat{\delta})$

Here, we have a single df to use for confidence intervals:

Degrees of Freedom $(n_B - 1) + (n_A - 1)$

Key points:

Follow the sample sizes for **each treatment level**

Each is used 2x

We will always use the **pooled** $s^2_{\hat{\mu}}$ in this class because of limitations of the lm() and lmer() functions

Estimated Standard Errors for each design

		Sit		Stand		Sit	# people	# measures	#EU
3)	Jill	X T1 X T2	Bob	X T1 X T2	Amy	X T1 X T2	40	80	40

Which equation for s^2 and SED?

Which components will **tend to be** different from Design 1 / 2 ?

* our data will be different, so all estimates will be different *

$$\hat{\delta}_{B-A} = \hat{\mu}_B - \hat{\mu}_A \qquad s^2_{\hat{\mu}} = \frac{\sum (\hat{\mu}_{Bj} - \hat{\mu}_B)^2 + \sum (\hat{\mu}_{Aj} - \hat{\mu}_A)^2}{(n_B - 1) + (n_A - 1)}$$

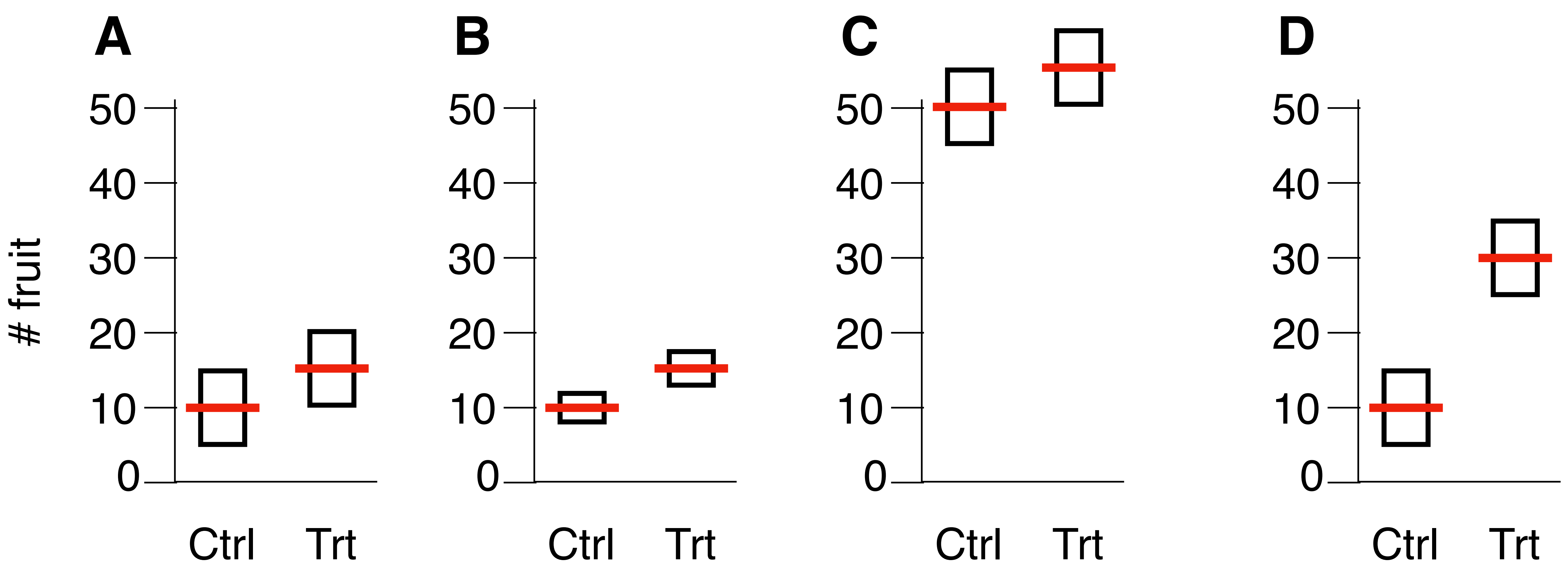
SED = $\sqrt{\dfrac{s^2_{\hat{\mu}}}{n_B} + \dfrac{s^2_{\hat{\mu}}}{n_A}}$

$s^2_{\hat{\mu}}$ will **tend to be** smaller because of less measurement error

n_B and n_A will be smaller (or equal) because of costs

If so, SED *might be* larger and DF would be smaller

$n_c = n_t = 32$



Which effect is largest?

Which effect is most important?

Which effect is most significant?

$\hat{\delta}$	5	5	5	20
s_{pooled}	8	2	8	8
SED	2	0.5	2	2
$\frac{\hat{\delta}}{\hat{\mu}_c}$	5/10=0.5	0.5	5/50=0.1	2
$\frac{\hat{\delta}}{SED}$	5/2=2.5	5/0.5=10	2.5	10

Hypothesis tests deal with significance, not importance

Hypothesis testing

Unlike confidence intervals, do not report effect sizes

Instead, report **evidence** or a **decision** about whether an effect could be 0

T-test:

1) Calculate t-statistic: $\frac{\hat{\delta}}{SED}$

2) Calculate **p-value** from T-distribution with *df*

```
2*pt(t,df,lower.tail=F)
```

Outcomes: Decision about plausibility of H_0

1) Weigh evidence

Is **p** small?

The smaller **p**, the stronger the evidence that $\delta \neq 0$

Report: more/less significant

p-value is one piece of evidence

weigh this with effect size, plausibility, other data

2) Decide Yes/No

Determine consequences of being wrong

Choose a threshold α

If $p < \alpha$, declare **significant**, state $\delta \neq 0$

If $p > \alpha$, declare **not significant**, state δ **could be** 0

Don't report the p-value itself!

Hypothesis testing

Determine consequences of being wrong

Table of outcomes

		Fail to Reject	**Reject**
		Declare δ may be 0	Declare $\delta \neq 0$
TRUE	$\delta = 0$	+	X - False Positive
FALSE	$\delta \neq 0$	X - False Negative	+

α = Probability of **False Positive (**Reject when $\delta = 0$)

If $\delta = 0$, and our $p < \alpha$, we'll make a False Positive mistake

Probability of this is α

β = Probability of **False Negative** (Accept when $\delta \neq 0$)

Lost opportunities

If $\delta \neq 0$, and our $p > \alpha$, we'll make a False Negative mistake

$1 - \beta$ = **Power** (Reject when $\delta \neq 0$)

Power: Change of declaring significant when $\delta \neq 0$

Goal: Power > 80%

What determines the Power of an experiment?

Declare significant if $p < \alpha$

What goes into p?

2*pt(t,df,lower.tail=F)

$$t = \frac{\hat{\delta}}{SED}$$

TRUE effect size δ

$$\sqrt{\frac{s^2_{pooled}}{n_B} + \frac{s^2_{pooled}}{n_A}}$$

$$\sigma^2_y = \sigma^2_\mu + \sigma^2_m$$

Sample size

df

Denominator of s^2_{pooled} $(n_A - 1) + (n_B - 1)$

What controls α?

You choose α!

Higher α -> higher power

But also greater chance of a False Positive

Calculating Power

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,
             power = NULL,
             type = c("two.sample", "one.sample", "paired"),
             alternative = c("two.sided", "one.sided"),
             strict = FALSE, tol = .Machine$double.eps^0.25)
```

n = # samples **per treatment**

delta = **TRUE** effect size

sd = **TRUE** standard deviation of observations

sig.level = α

power = $1 - \beta$

Choose 1 of these to set to NULL

R will calculate its value

Need to guess at **delta** and **sd**

Questions:

What happens to **Power** when you **increase** each of the other parameters?

List 4 ways in **increase Power** in an experiment

Calculating Power

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,
             power = NULL,
             type = c("two.sample", "one.sample", "paired"),
             alternative = c("two.sided", "one.sided"),
             strict = FALSE, tol = .Machine$double.eps^0.25)
```

n = # samples **per treatment**

delta = **TRUE** effect size

sd = **TRUE** standard deviation of observations

sig.level = α

power = $1 - \beta$

Other options:

type: two.sample = Replicate Level

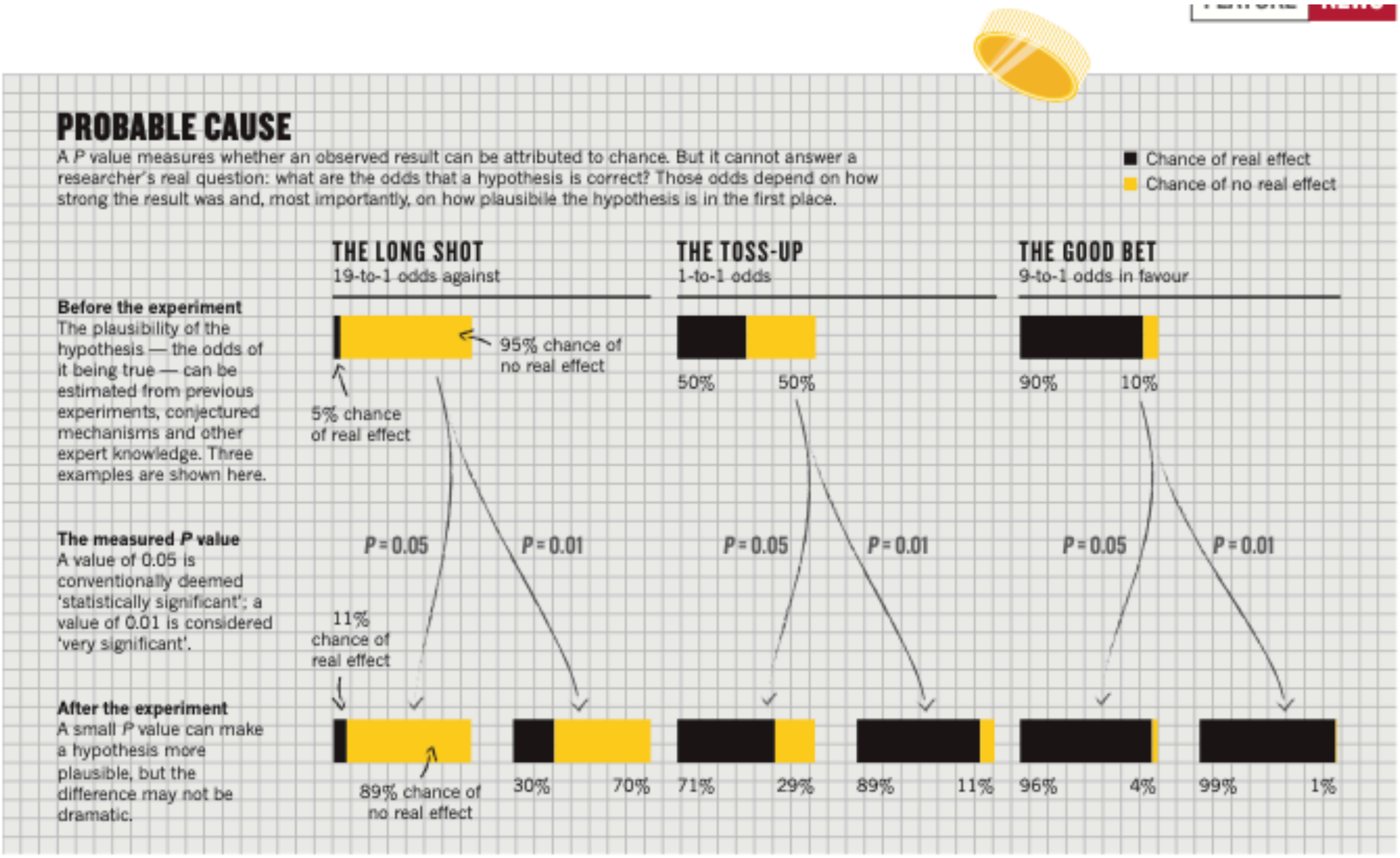
paired = Replicate Effect

one.sample = Test if $\mu_A = 0$

alternative: two.sided: test if $\delta \neq 0$

one.sided: test if $\delta > 0$

Statistical Errors paper



Key points

small p-value from a implausible treatment is not strong evidence

small p-value from an experiment with low power won't replicate

you can get a small p-value with a meaningless effect if your experiment is large

If your experiment is small and your p-value is small, your effect size is probably over-estimated

Rules for making Design Table

Include all variable necessary to describe the experiment

- Treatments: Variables we want to study
- Response: One Variable, always numeric
- Design
  - EU of the Treatment variable(s)
  - any Replicate and Replicate:Treatment
  - Variable with a unique level for each observation (Response)
  - Any other variable to describe the experiment

Check variable relationships: nested, aliased and crossed

- EU Variable must be **nested** in the Treatment variable
- If two variables are **aliased**, keep only 1 of them
- If two variables are **crossed**, keep only both

Relationships among variables

nested many:one Keep both, if 1 first is random, so is second

Experiment 2

Person	Person:Trial
Jill	Jill:T1
	Jill:T2
Bob	Bob:T1
	Bob:T2
Amy	Amy:T1
	Amy:T2

is_nested(Person~Person:Trial)

aliased one:one Keep one, particularly EU

Experiment 2

Person:Trial	Person:Posture
Jill:T1	Jill:Sit
Jill:T2	Jill:Stand
Bob:T1	Bob:Stand
Bob:T2	Bob:Sit
Amy:T1	Amy:Sit
Amy:T2	Amy:Stand

is_aliased(A~B)

crossed many:many Keep both

Experiment 2

Experiment 3

Posture	Trial
Sit	T1
Stand	T2

is_crossed(A~B)

Rules for making Design Table

Include all variable necessary to describe the experiment

- Treatments: Variables we want to study
- Response: One Variable, always numeric
- Design
  - EU of the Treatment variable(s)
  - any Replicate and Replicate:Treatment
  - Variable with a unique level for each observation (Response)
  - Any other variable to describe the experiment

Check variable relationships: nested, aliased and crossed

- EU Variable must be **nested** in the Treatment variable
  - Label as EU:Treatment
- If two variables are **aliased**, keep only 1 of them
  - If one is an EU, keep that one!
- If two variables are **crossed**, keep both
  - Treatments are crossed with their Replication variable

3)

	Sit			Stand			Sit		# people	# measures	#EU
Jill	X	T1	Bob	X	T1	Amy	X	T1	40	80	40
	X	T2		X	T2		X	T2			

Person	Posture	Pulse	Trial	Person:Trial
Jill	Sit	60	T1	Jill:T1
Jill	Sit	64	T2	Jill:T2
Bob	Stand	72	T1	Bob:T1
Bob	Stand	68	T2	Bob:T2
Amy	Sit	106	T1	Amy:T1
Amy	Sit	112	T2	Amy:T2
⋮				

Structure	Variable	Type	#levels	Replicate	EU
Treatment	Posture	Cat	2	None	Person
Design	Person	Cat	40		
	Trial	Cat	2		
	Person:Trial	Cat	80		
Response	Pulse	Num	80		