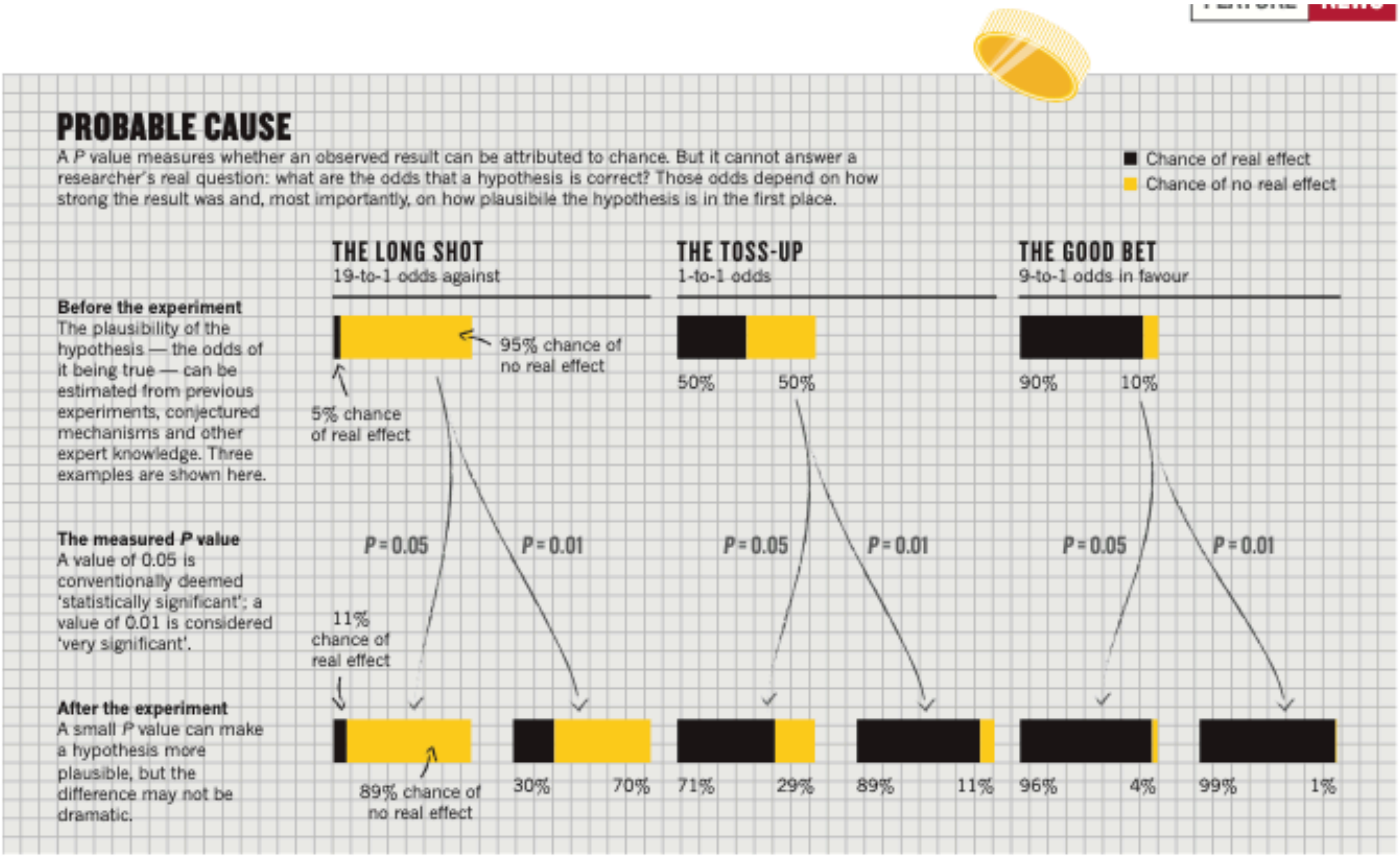
Statistical Errors paper



A p-value is an answer to the question:

“Is it plausible that the TRUE effect was 0?”

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

Power: Probability of detecting an effect when it is real

Detection threshold: Declare significant if $p < \alpha$

What determines the Power of an experiment?

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 = No replication

paired = Replicated treatment effects

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

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

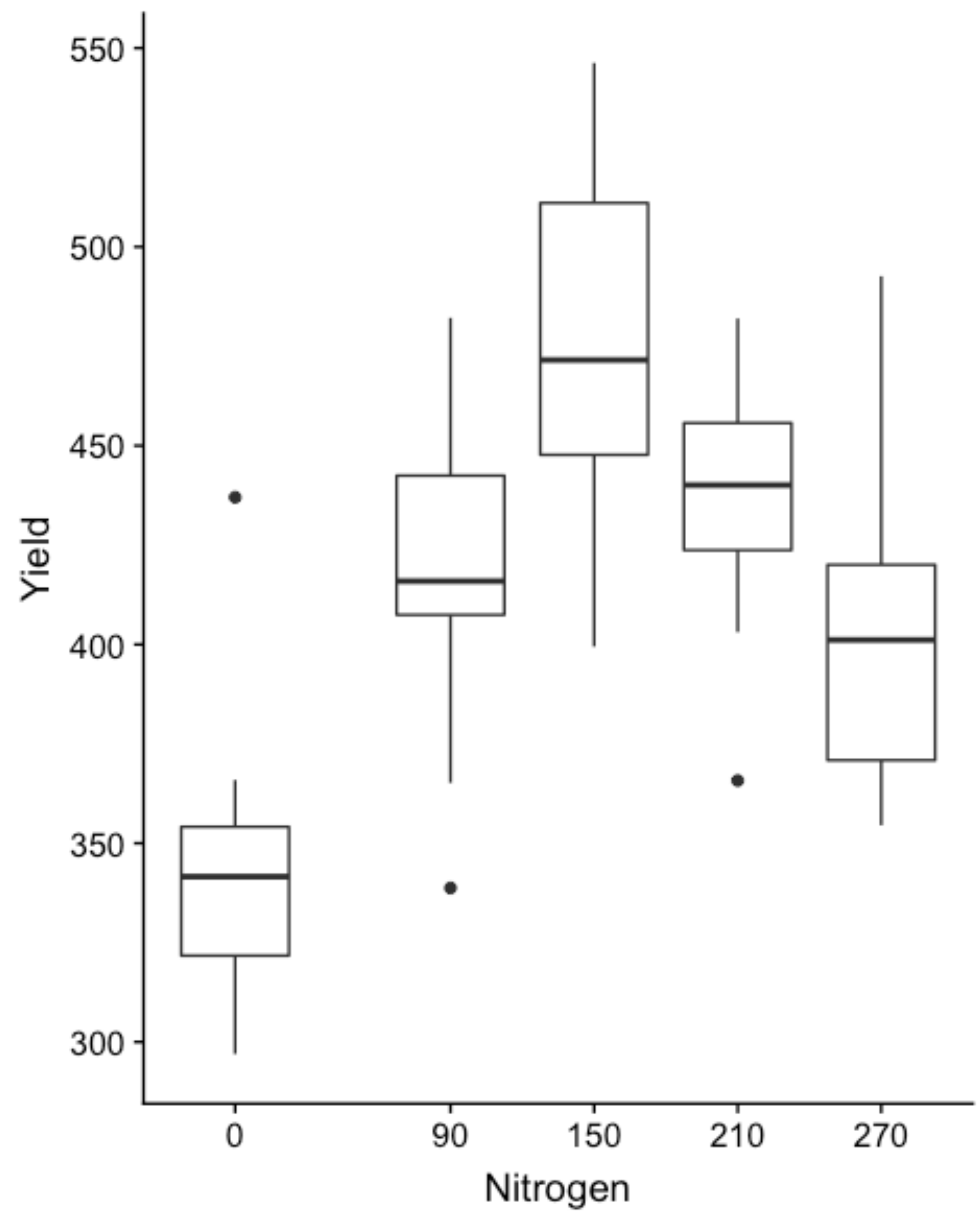
one.sided: test if $\delta > 0$

An experiment was run to evaluate effects of increased nitrogen fertilization on tuber yield of frying potatoes

5 nitrogen regimes (applied to plots):
0, 90, 150, 210, 270 lbs / acre at emergence

10 reps / treatment combination

Response: total yield per plot



Structure	Variable	Type	#levels	Replicate	Experimental Unit
Treatment	Nitrogen	Categorical or Numeric	5	None	Plot
Design	Plot	Categorical	50		
Response	Yield	Numeric	50		

What questions should we address with these data?

“Questions” should be phrased around treatment effects (δ_{E-B}, etc)

Is +Nitrogen better than 0-Nitrogen?

Which [Nit] change Yield?

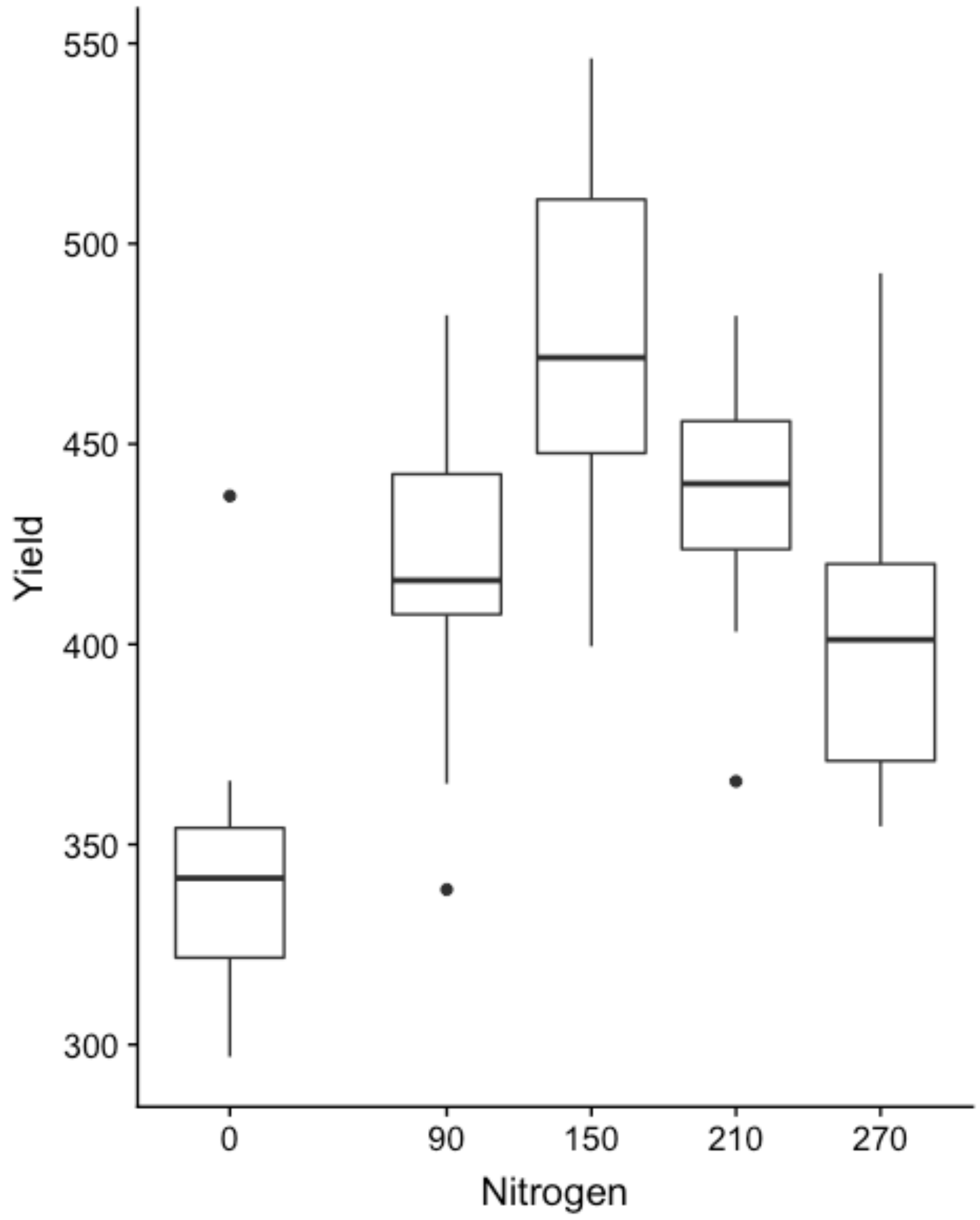
What [Nit] increases Yield the most relative to 0-Nitrogen

An experiment was run to evaluate effects of increased nitrogen fertilization on tuber yield of frying potatoes

5 nitrogen regimes (applied to plots): 0, 90, 150, 210, 270 lbs / acre at emergence

10 reps / treatment combination

Response: total yield per plot



Reporting $\hat{\delta}_{90-0}$:

Estimate $\hat{\delta}_{90-0} = \hat{\mu}_{90} - \hat{\mu}_0$

Calculate SED = $\sqrt{\dfrac{s_p^2}{n_{90}} + \dfrac{s_p^2}{n_0}}$

Report CI: $\hat{\delta}_{90-0} \pm t_c \times SED$

Note: s^2_{pooled} uses replicates from all 5 treatments

df = 5*(10-1)

Reporting $\hat{\delta}_{150-0}$:

Same …

Reporting $\hat{\delta}_{150-90}$:

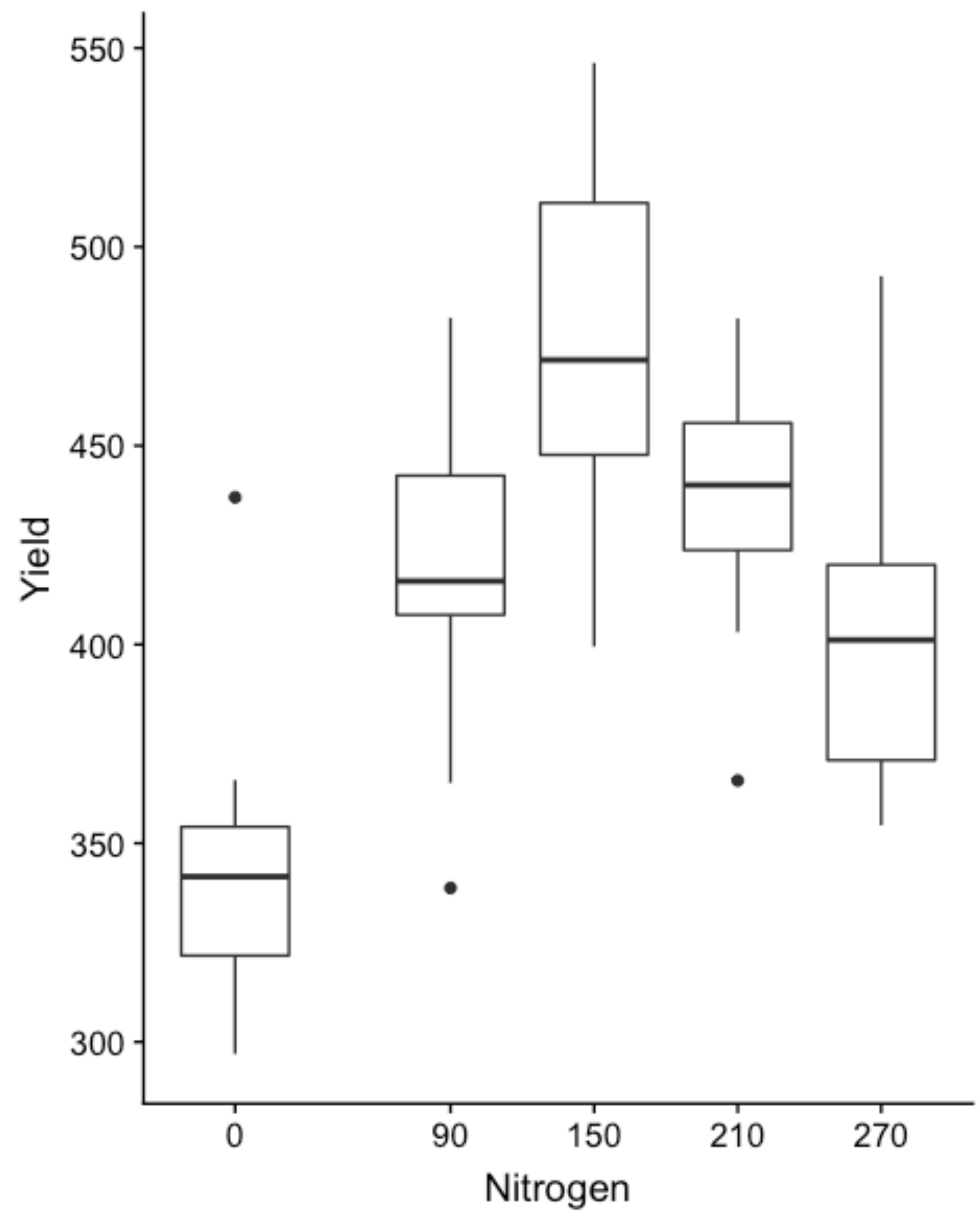
Same …

An experiment was run to evaluate effects of increased nitrogen fertilization on tuber yield of frying potatoes

5 nitrogen regimes (applied to plots):
0, 90, 150, 210, 270 lbs / acre at emergence

10 reps / treatment combination

Response: total yield per plot



What is the maximum improvement we could get?

Answer: What is the effect of the best level of Nitrogen?

Report $\hat{\delta}_{150-0} \pm t_c \times SED$

Can any addition of Nitrogen actually increase yield?

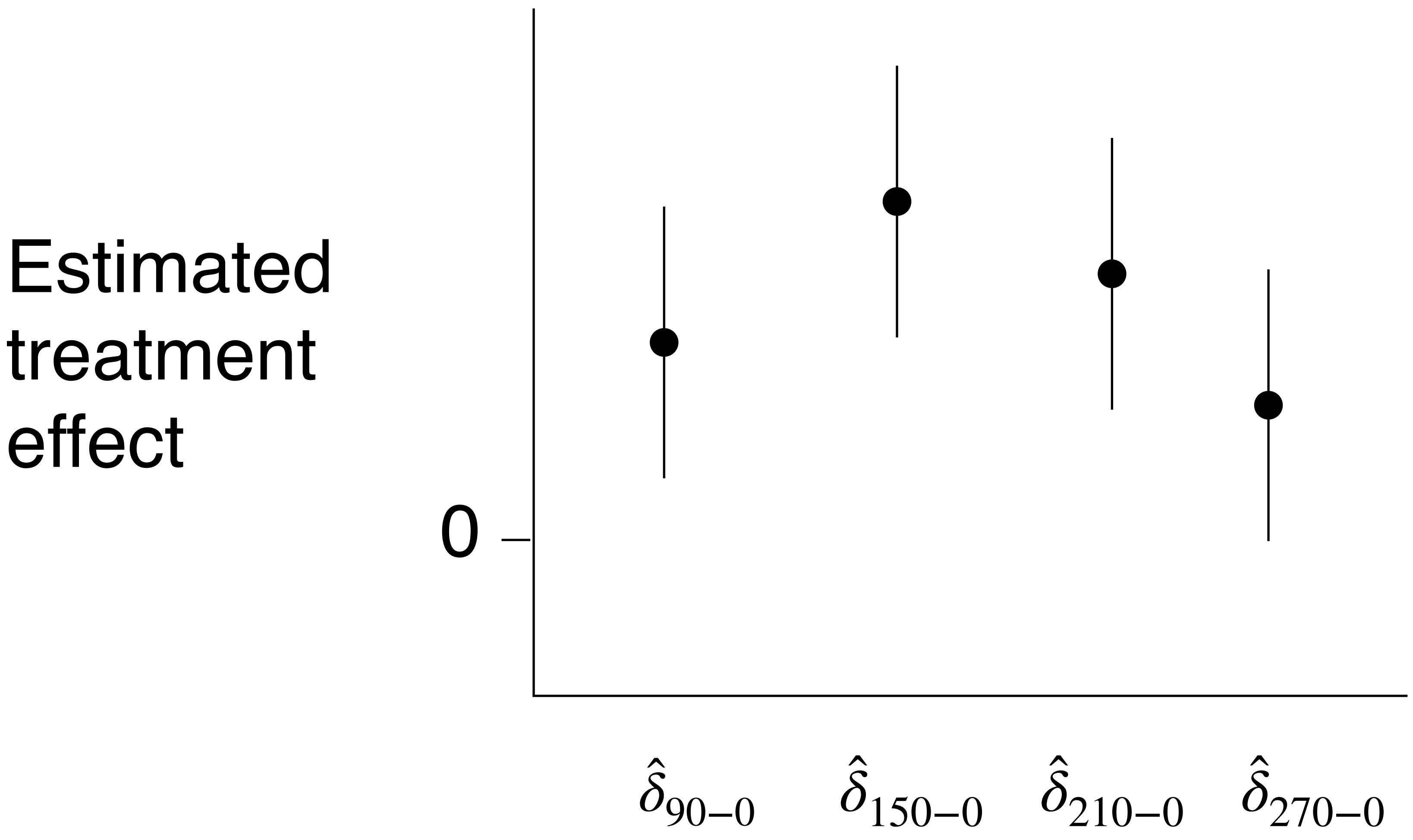
Answer: Power is highest when δ is biggest

Run T-test for $\delta_{150-0} = 0$

Both of these answers are misleading because we chose to run the statistics **because** the **estimated yield** for N-150 was highest

Once you look at the data, neither CIs nor p-values are valid

With 4 **new** treatments each compared to the control we are making 4 estimates



For each CI, there's a 5% chance that the TRUE effect is outside the interval

There is a 15% chance that the TRUE effect is outside **at least one** of these intervals

What is the maximum improvement we could get?

There's a good chance the biggest estimated effect was over-estimated

We're "safe" if we can ensure all CIs include their TRUE values

Strategy: Adjust CIs so that the chance that **any** true effect is outside of the interval is $100\alpha\,\%$

CI: $\hat{\delta}_{i-0} \pm t_c^D \times SED$

$t_{\alpha,df}^{D(k)}$ comes from the Dunnett distribution

k: # **new** treatments (excluding control)

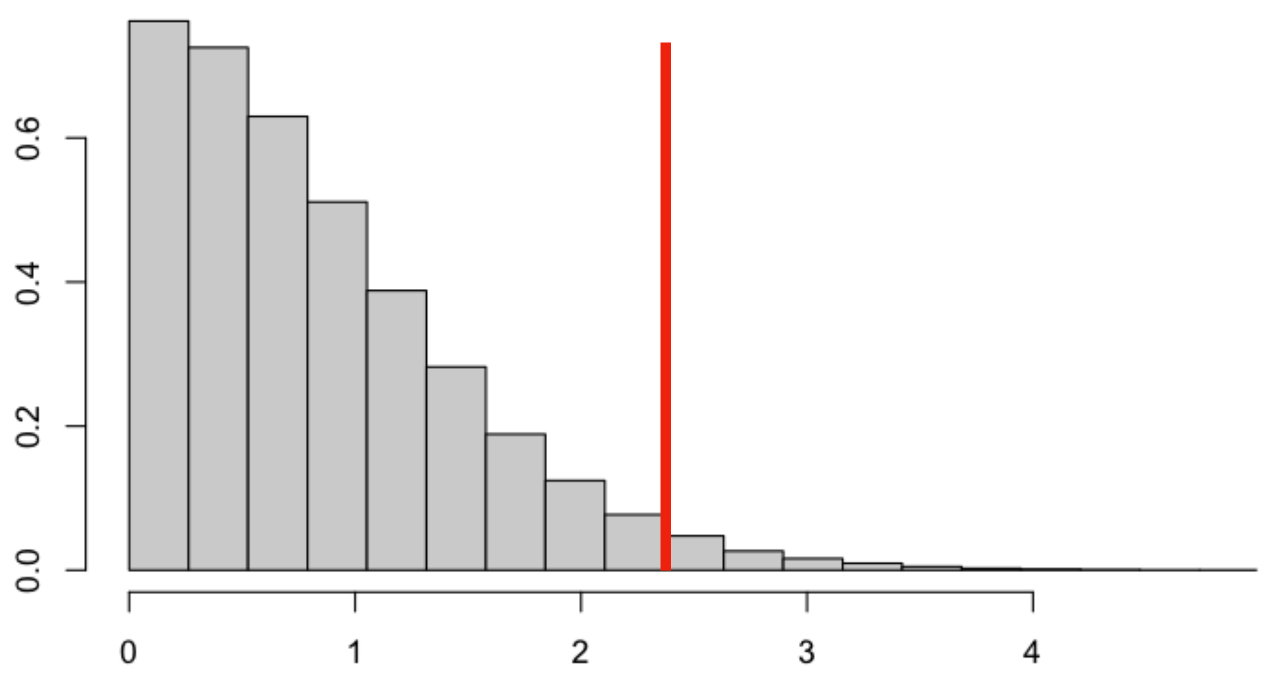
α: False Positive rate

df: Degrees of freedom from all treatments

Bigger value than t_c

Accounting for **multiple comparisons**

T-distribution



$$\frac{|\hat{\delta} - \delta|}{\text{SED}}$$

Actual error

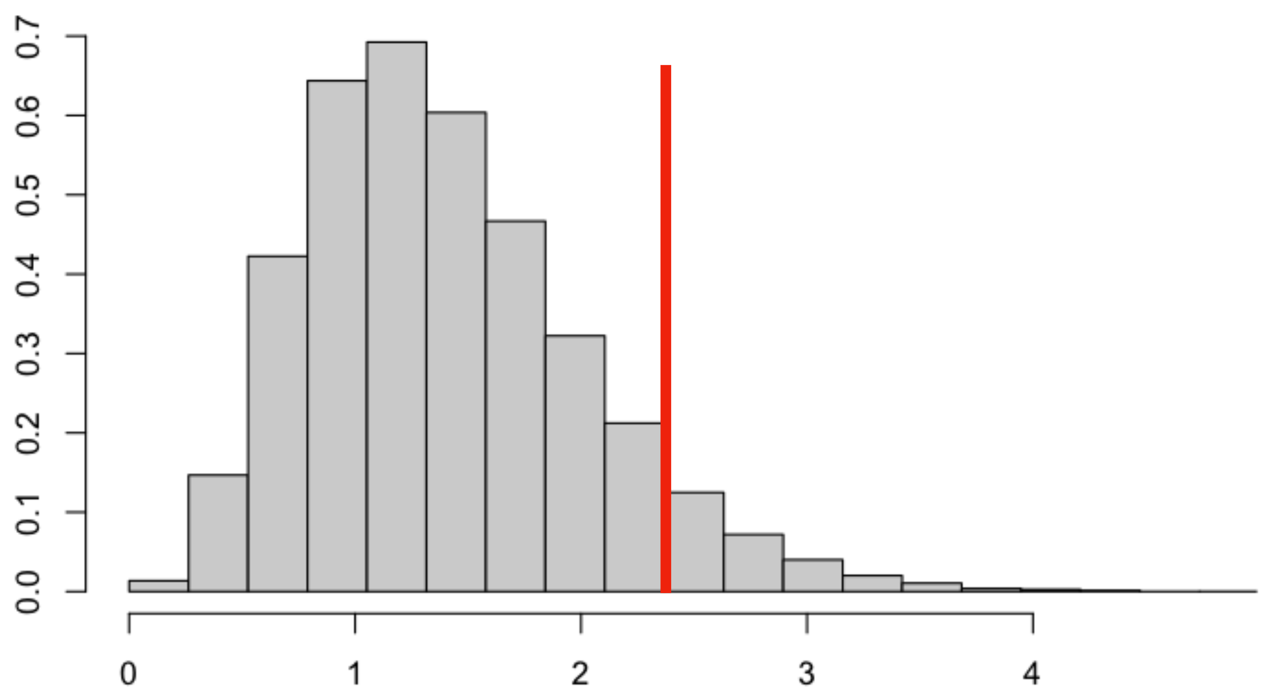
Estimated average error

Distribution of Normalized errors

How much bigger than SED could my actual error have been?

*Valid for a single treatment effect

Dunnett(4) distribution



Estimate: $\hat{\delta}_{90-0}$, $\hat{\delta}_{150-0}$, $\hat{\delta}_{210-0}$, $\hat{\delta}_{270-0}$

$$\frac{max|\hat{\delta}_i - \delta_i|}{\text{SED}}$$

Actual size of **biggest** error

Estimated average error

Distribution of Biggest Normalized errors

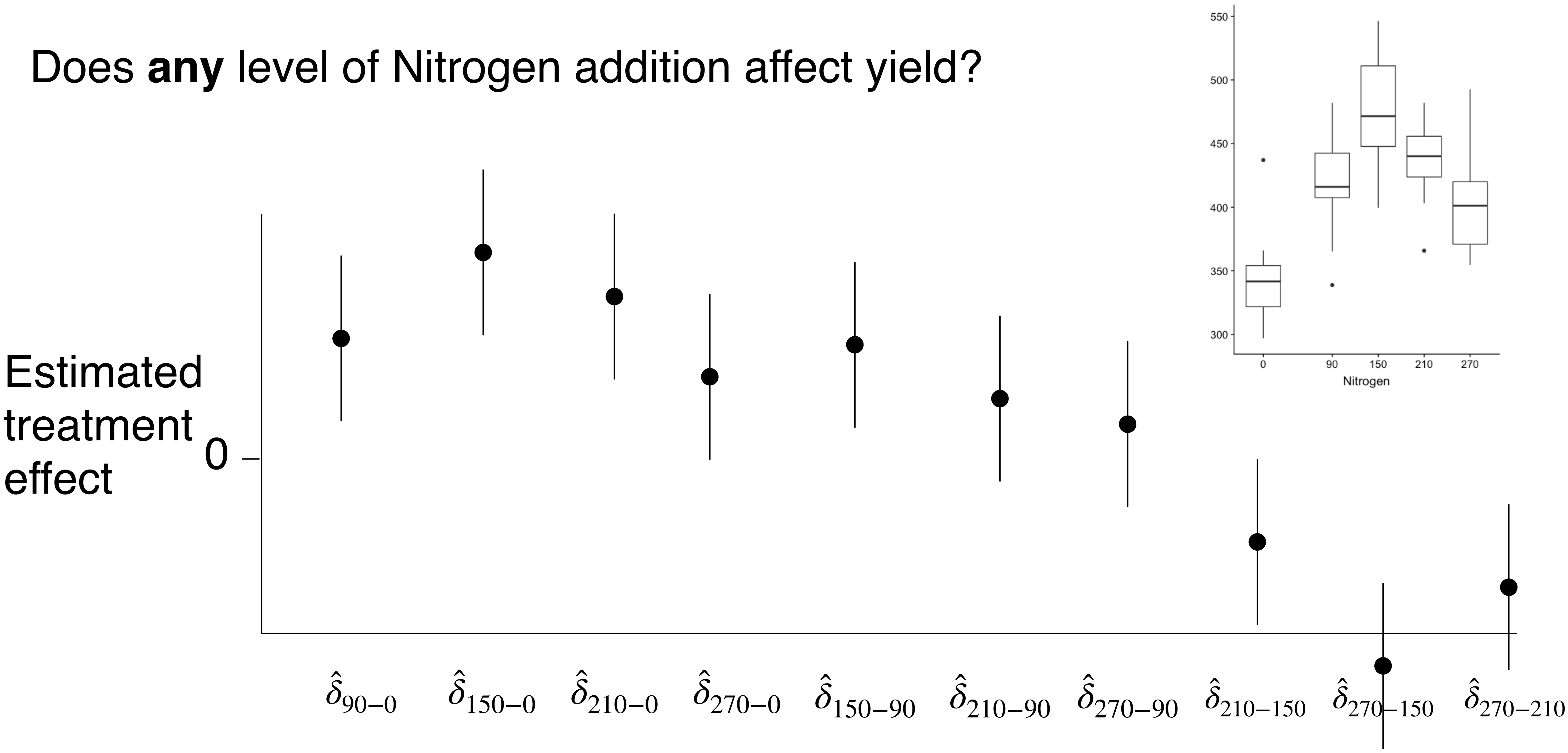
Can any addition of Nitrogen actually increase yield?

If we run a T-test for $\delta_{150-0} = 0$, the p-value will be **too small**

Instead calculate p-value from the Dunnett(4) distribution

Corrected p-value: Probability of **the biggest observed Normalized effect** being this large if all 4 treatments had no effect

Does **any** level of Nitrogen addition affect yield?



With 5 treatment levels, we can make 10 pairwise comparisons

At least 1 CI won't include the TRUE effect ~27% of the time

We'd conclude that at least 2 levels differ ~27% of the time even if Nitrogen had no effect at all

Solution:

CI: $\hat{\delta}_{i-0} \pm t_c^T \times SED$

$t_{\alpha,df}^{T(k)}$ comes from the Tukey distribution

k: # **total** treatments

α: False Positive rate

df: Degrees of freedom from all treatments

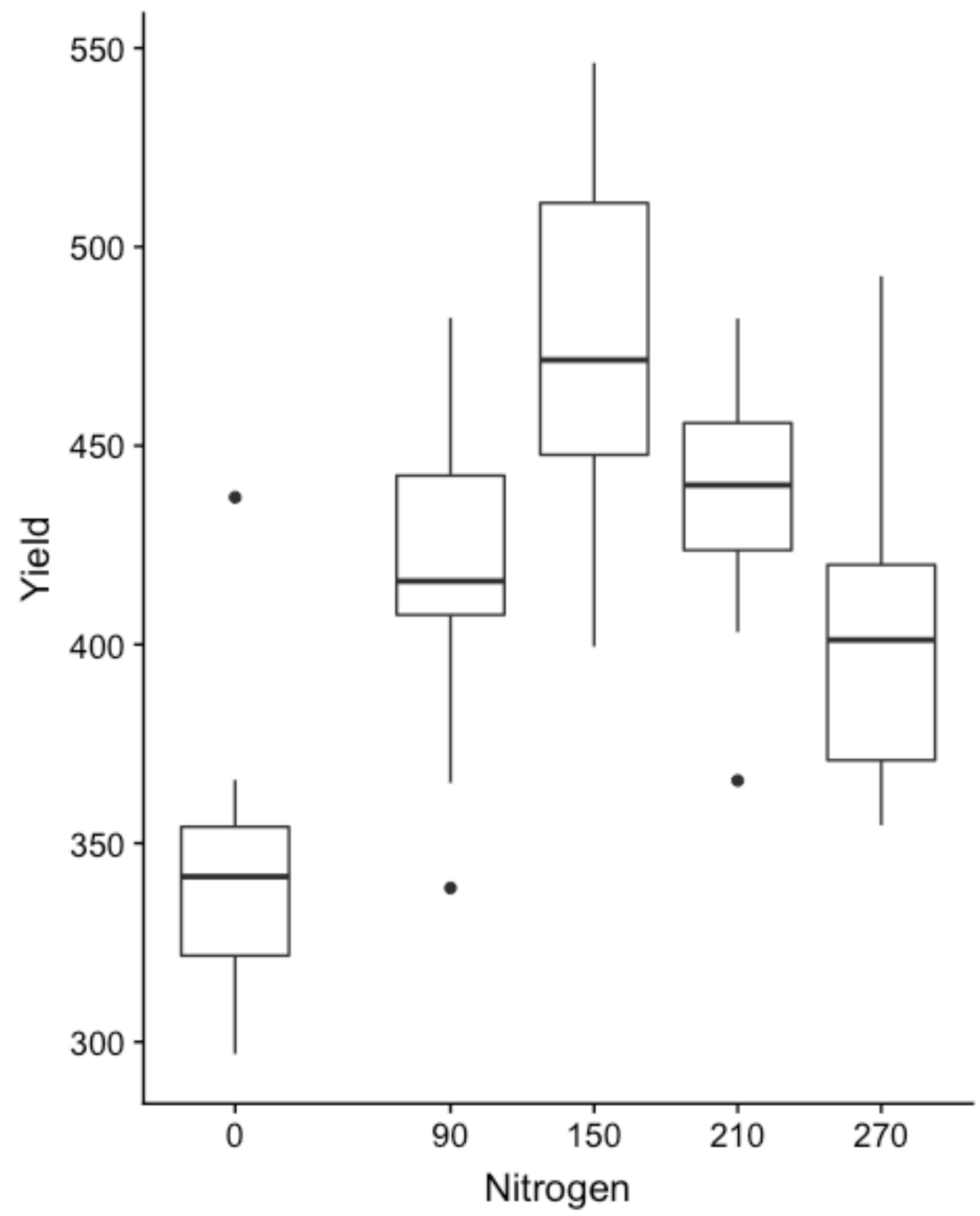
Bigger value than t_c from either T or Dunnett

An experiment was run to evaluate effects of increased nitrogen fertilization on tuber yield of frying potatoes

5 nitrogen regimes (applied to plots): 0, 90, 150, 210, 270 lbs / acre at emergence

10 reps / treatment combination

Response: total yield per plot



If you are specifically asked about the effect of N=90 vs N=0:

Use the T-distribution

Specify: emmeans(…, at=list(Nitrogen=c(0,90))

If you are interested in which (if any) are different from the control (N=0)

Use the Dunnett distribution

Cannot compare the new treatments

Remember: Not significant ≠ No effect

Specify: contrast(…,method = ‘trt.vs.ctrl’,ref=‘0’)

If you are interested in which (if any) are different from any other

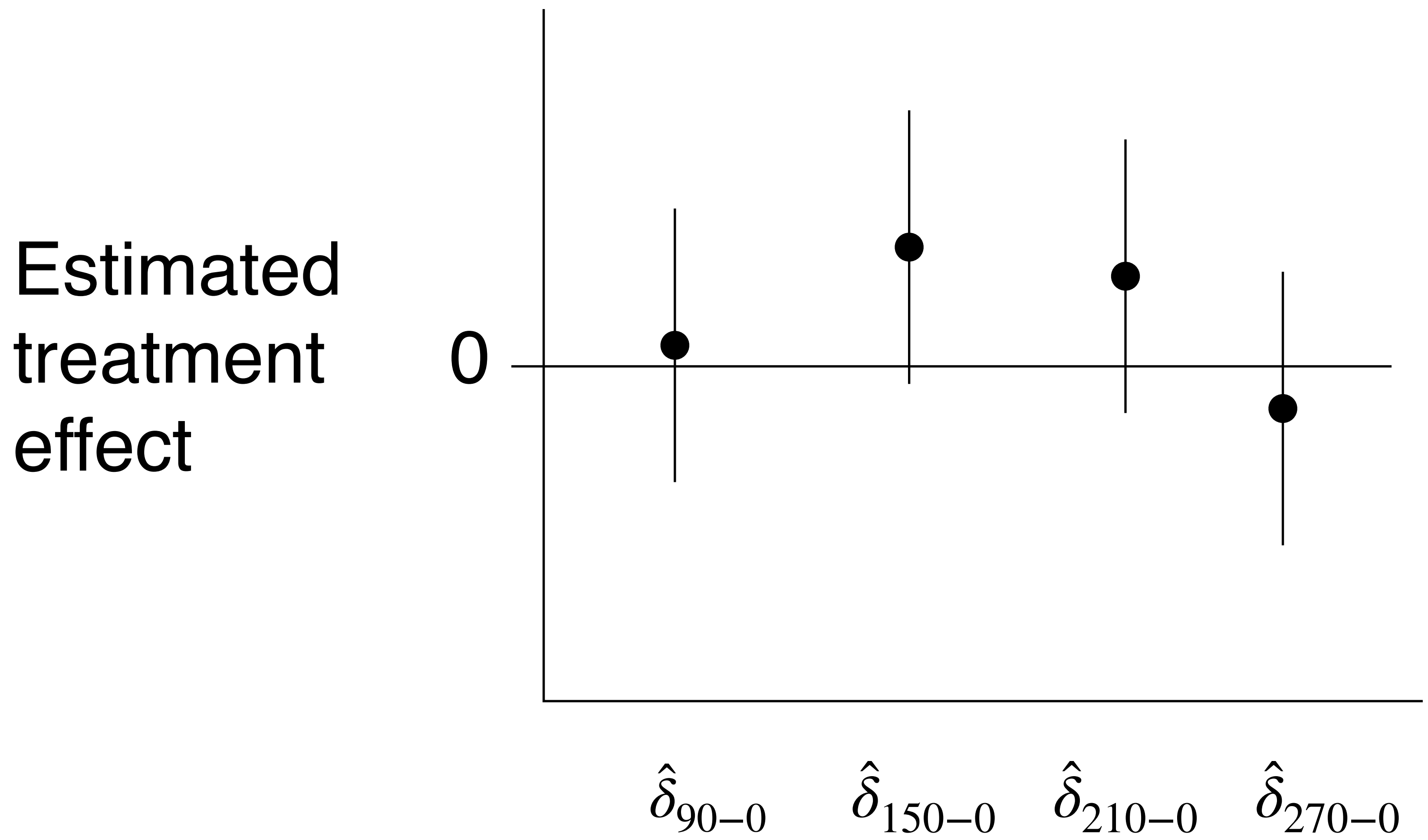
Use the Tukey distribution

Most common situation

Specify: contrast(…,method = ‘pairwise’)

What happens if you don't account for multiple comparisons?

Say Nitrogen actually had ZERO effect on yield…



"Normal CI's"

5% of the time the CI for $\hat{\delta}_{150-0}$ would not cross zero (False Positive)

15% of the time at least one of the 4 CIs would not cross zero

15% of the time you would conclude "Nitrogen has an effect"

Our confidence in the $\hat{\delta}_{150-0}$ doesn't change with the number of levels of Nitrogen assayed

But if we pick out the **top effect(s)** or **significant effects** our confidence in their effect sizes is lower

Scenario: you're evaluating 10 fertilizer products from 10 companies

A) You will report back to each company the effect of their product

Distribution: T-distribution

B) You are reporting to a store which products are worth putting on the shelves

Distribution:

Dunnett(10) if all that improve yield over no fertilizer will be sold

Tukey(11) if only the best few will be sold

What can go wrong?

~~Estimates are biased~~

CI too small

CI too big

Assumptions for calculating Confidence Intervals

1) EU are independent

Count n for SEM and *df*

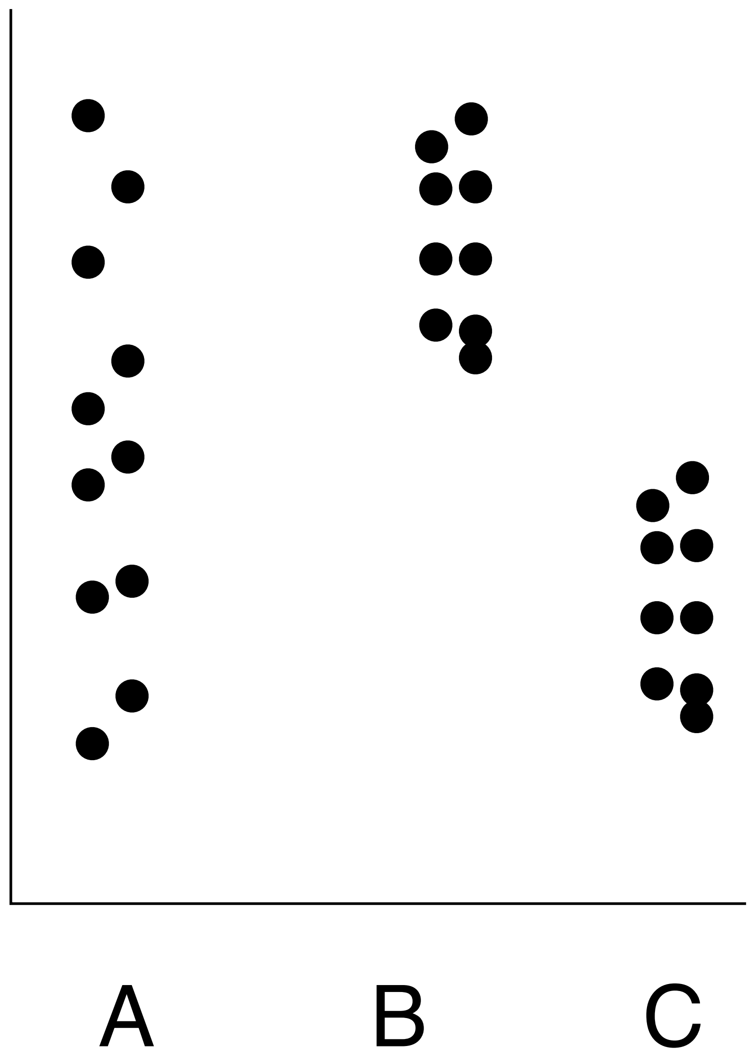
2) $\sigma^2_{\mu_i}$ and σ^2_m are the same across groups

Pooling deviations to calculate s^2_{pooled}, maximizing *df*

3) μ_{ij} and ϵ_{ij} are Normally distributed

T-distributions, Confidence Intervals and p-values

What can go wrong?



s^2_{pooled} is a weighted average of s_i^2

SEs for treatment means:

$$SEM_A = \sqrt{\frac{s_p^2}{n_A}}$$ too small

$$SEM_B = \sqrt{\frac{s_p^2}{n_B}}$$ too large

CIs for treatment effects:

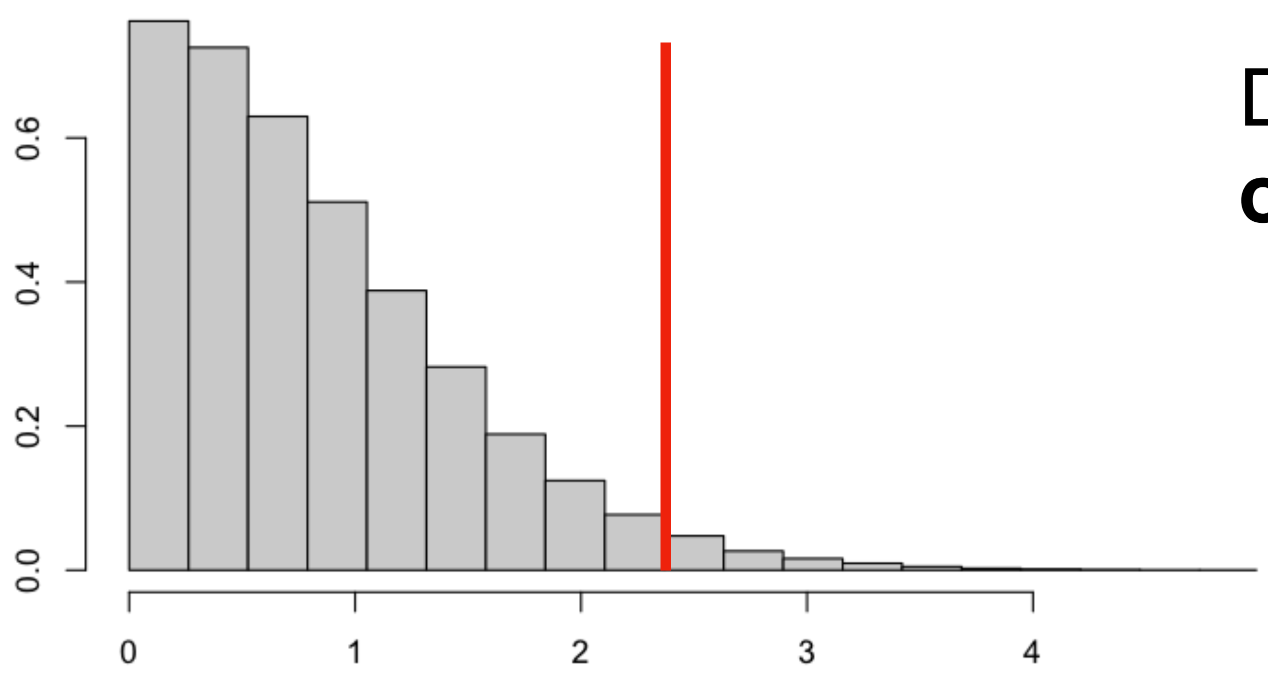
$\hat{\delta}_{B-A} \pm t_c \times SED$ too small

$\hat{\delta}_{C-A} \pm t_c \times SED$ too small

$\hat{\delta}_{C-B} \pm t_c \times SED$ too large

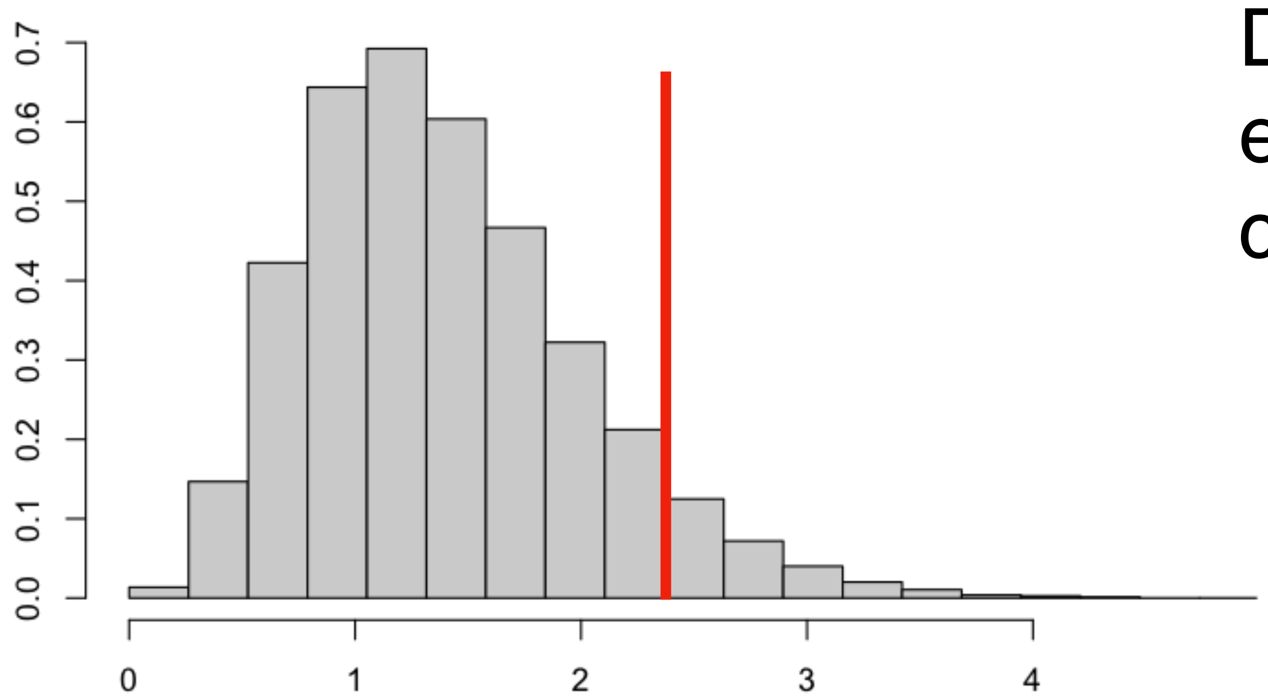
Accounting for **multiple comparisons**

T-distribution



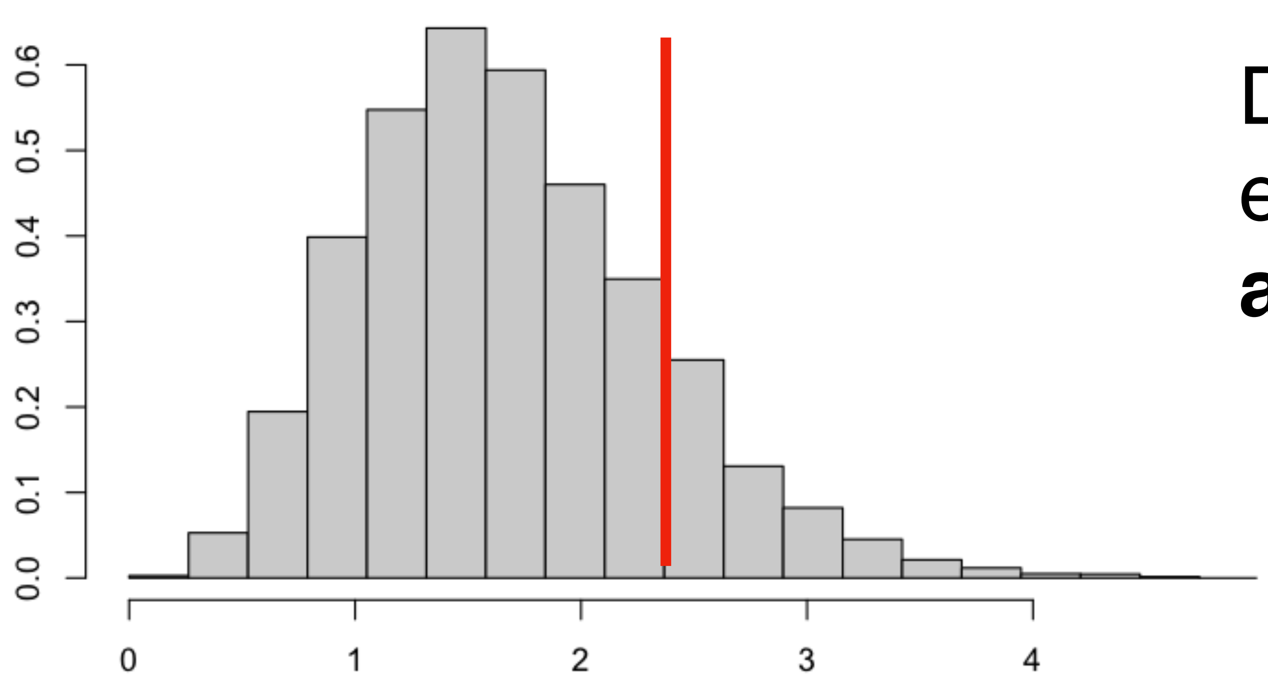
Distribution of Normalized errors for **one specific treatment effect**

Dunnett(4) distribution



Distribution of the biggest Normalized error among **4 new treatments** compared against a single control

Tukey(5) distribution



Distribution of the biggest Normalized error among **5 treatments compared in all pairwise combinations**