A Statistical Approach to Language Model Evaluations

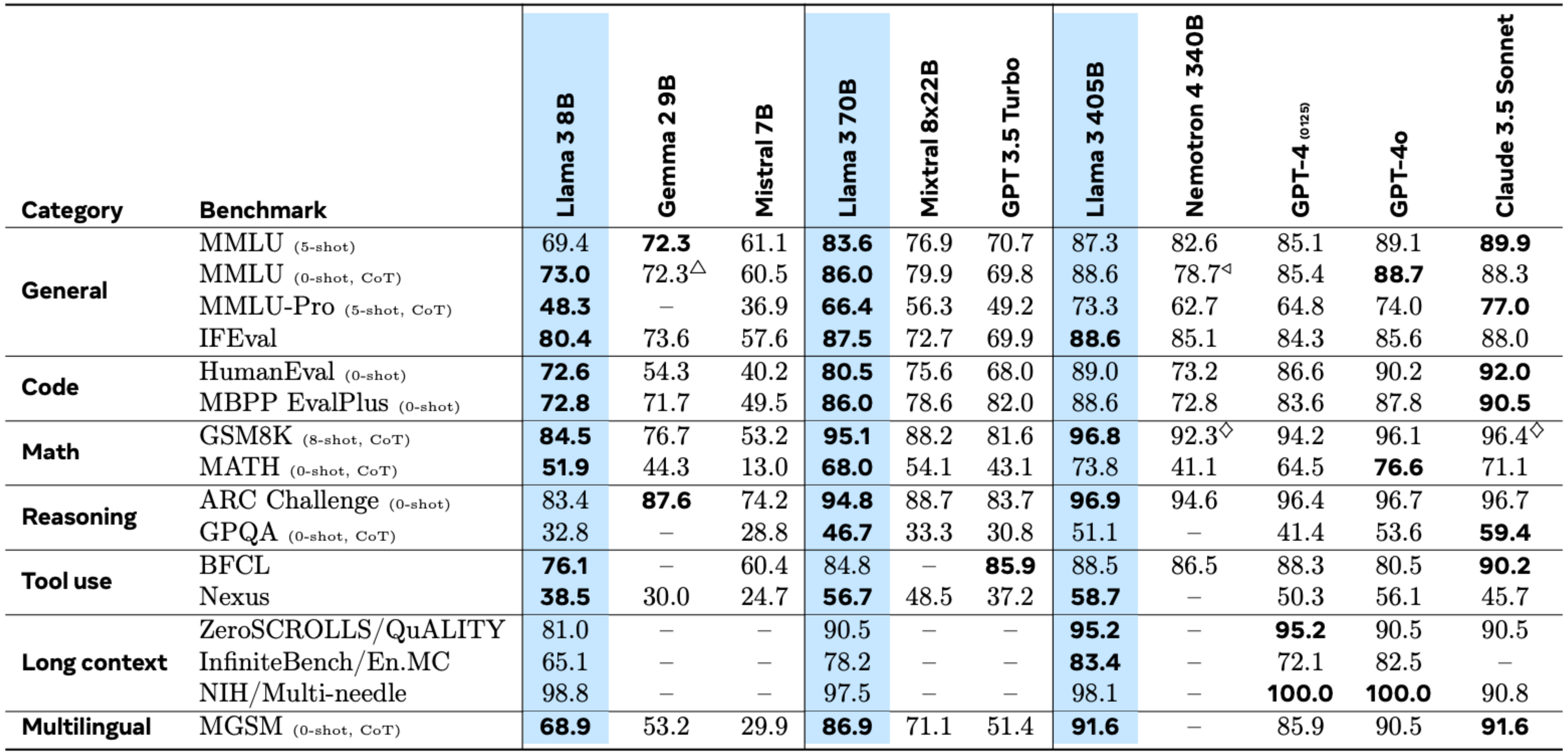
fzeng, 2024-11-18

Challenges with LLM Evals

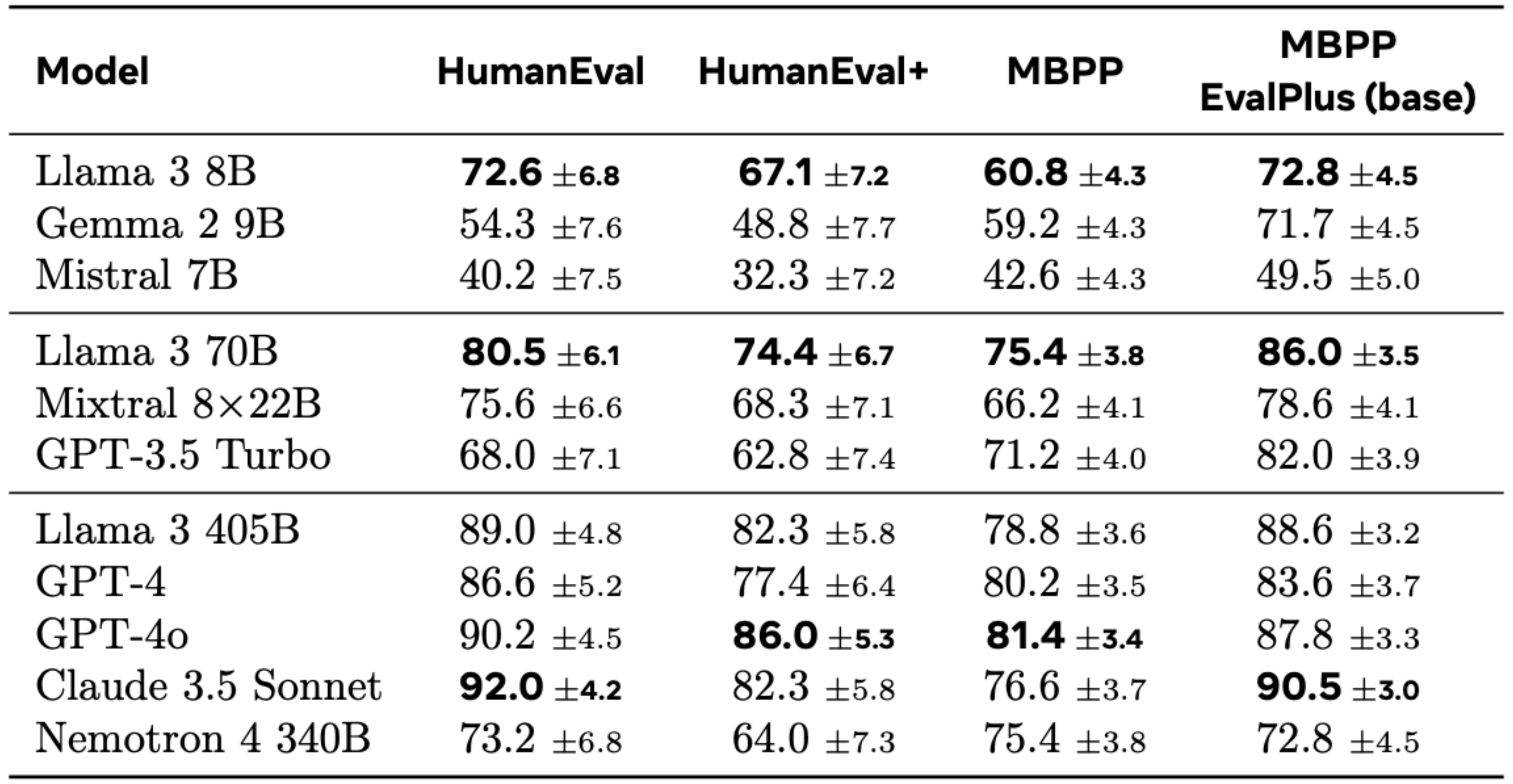
- LLM highly sensitive to prompts (Liang et al., 2022; Mizrahi et al., 2023; Scalar et al., 2023; Weber et al., 2023, Bsharat et al., 2023)
- Several widely used open-source LLMs extremely sensitive to subtle changes in prompt formatting in few-shot settings, with performance differences of up to 76 accuracy points (Scalar et al., 2023)
- Changing the options from (A) to (1) or changing the parentheses from (A) to [A], or adding an extra space between the option and the answer can lead to a ~5 percentage point change in accuracy on the evaluation (Anthropic)
- Tipping a language model 300K for a better solution" leads to increased capabilities (Bsharat et al., 2023)

We need a Science of Evals — Apollo Research

Papers often don’t report standard errors

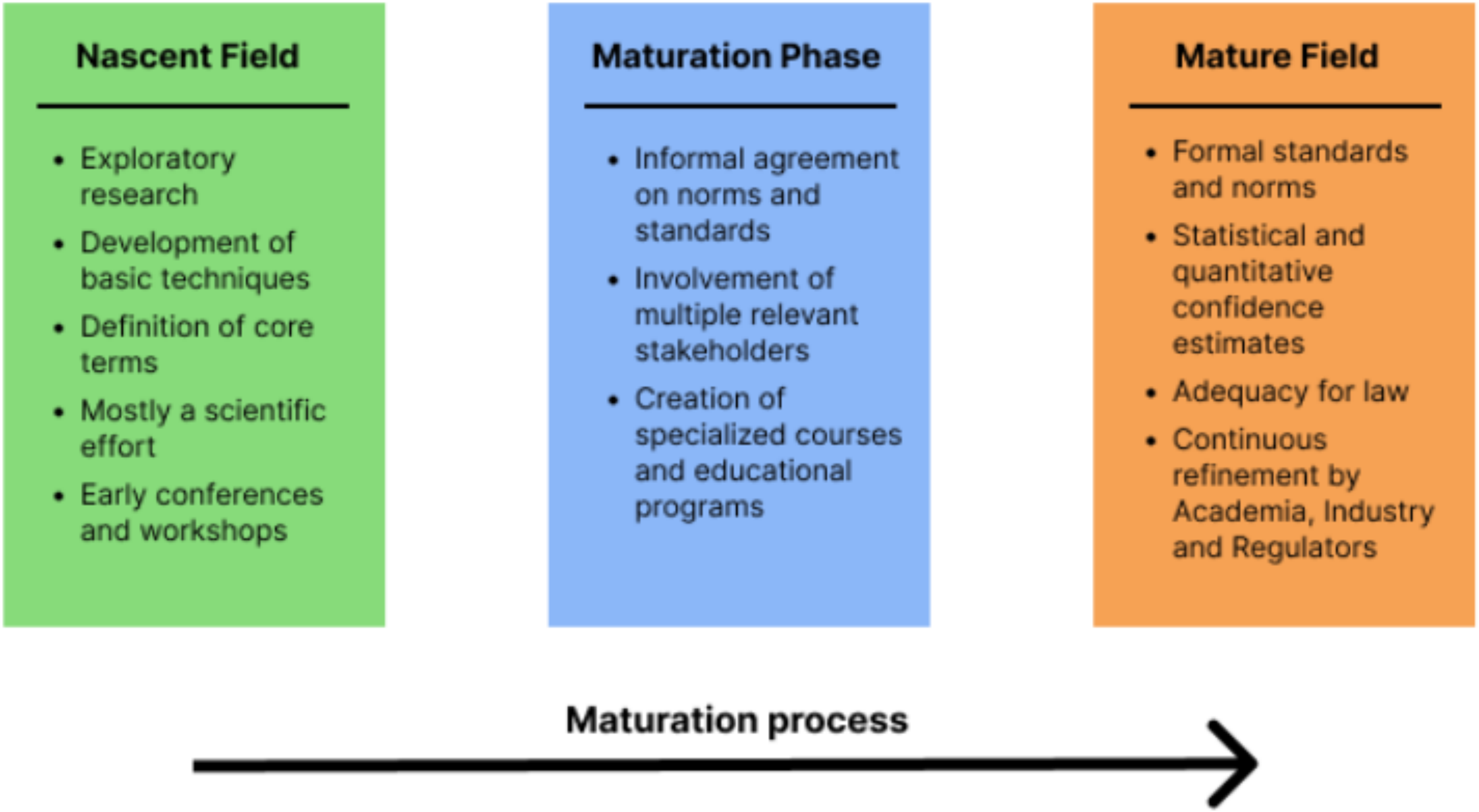


Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**



Model	HumanEval	HumanEval+	MBPP	MBPP EvalPlus (base)
Llama 3 8B	**72.6** ±6.8	**67.1** ±7.2	**60.8** ±4.3	**72.8** ±4.5
Gemma 2 9B	54.3 ±7.6	48.8 ±7.7	59.2 ±4.3	71.7 ±4.5
Mistral 7B	40.2 ±7.5	32.3 ±7.2	42.6 ±4.3	49.5 ±5.0
Llama 3 70B	**80.5** ±6.1	**74.4** ±6.7	**75.4** ±3.8	**86.0** ±3.5
Mixtral 8×22B	75.6 ±6.6	68.3 ±7.1	66.2 ±4.1	78.6 ±4.1
GPT-3.5 Turbo	68.0 ±7.1	62.8 ±7.4	71.2 ±4.0	82.0 ±3.9
Llama 3 405B	89.0 ±4.8	82.3 ±5.8	78.8 ±3.6	88.6 ±3.2
GPT-4	86.6 ±5.2	77.4 ±6.4	80.2 ±3.5	83.6 ±3.7
GPT-4o	90.2 ±4.5	**86.0** ±5.3	**81.4** ±3.4	87.8 ±3.3
Claude 3.5 Sonnet	**92.0** ±4.2	82.3 ±5.8	76.6 ±3.7	**90.5** ±3.0
Nemotron 4 340B	73.2 ±6.8	64.0 ±7.3	75.4 ±3.8	72.8 ±4.5

Science of Evals still young



Nascent Field

- Exploratory research
- Development of basic techniques
- Definition of core terms
- Mostly a scientific effort
- Early conferences and workshops

Maturation Phase

- Informal agreement on norms and standards
- Involvement of multiple relevant stakeholders
- Creation of specialized courses and educational programs

Mature Field

- Formal standards and norms
- Statistical and quantitative confidence estimates
- Adequacy for law
- Continuous refinement by Academia, Industry and Regulators

Maturation process

Paper recommendations

1. Computing standard errors of the mean using the Central Limit Theorem
2. When questions are drawn in related groups, computing clustered standard errors
3. Reducing variance by resampling answers and by analyzing next-token probabilities
4. When two models are being compared, conducting statistical inference on the question level paired differences, rather than the population-level summary statistics
5. Using power analysis to determine whether an eval (or a random subsample) is capable of testing a hypothesis of interest

Standard Errors of the Mean

Standard Error of the Mean

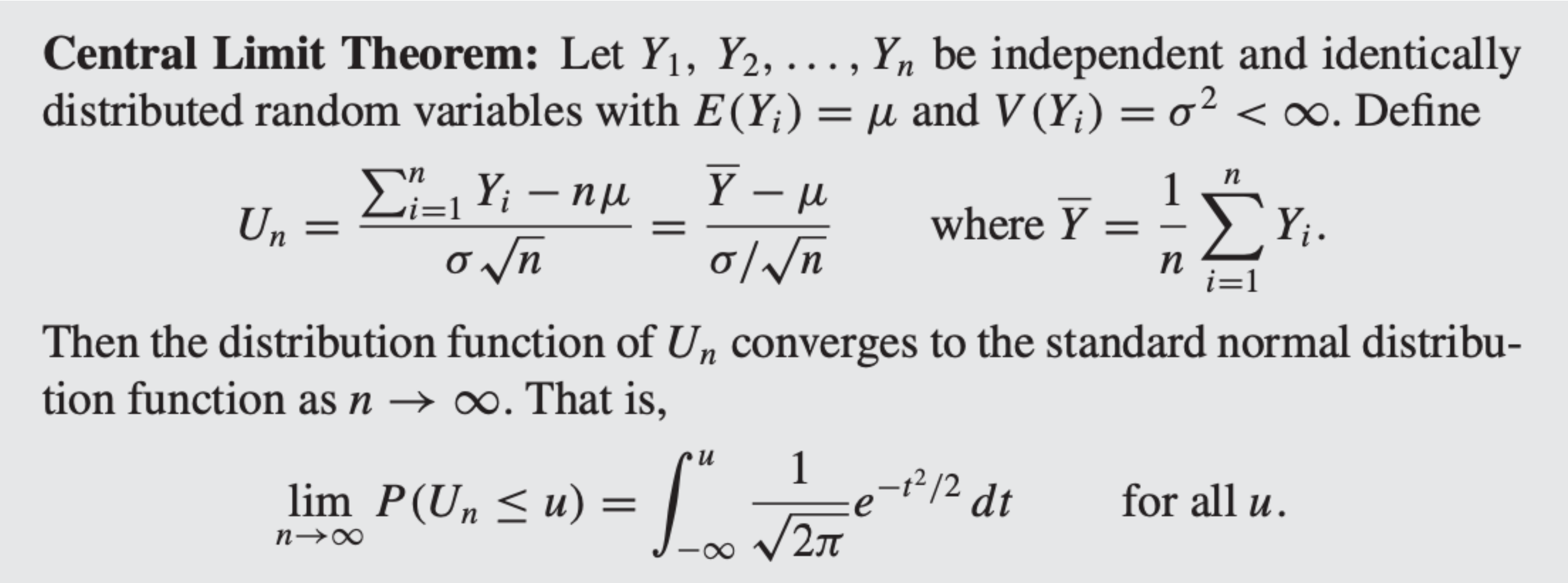
- Some notation:
- For some question i in the dataset,

$$\underbrace{s_i}_{\text{score of question i}} = \underbrace{x_i}_{\text{conditional mean on question i}} + \underbrace{\epsilon_i}_{\text{conditional variance on question i}}$$

- Can also talk about any question in the dataset unconditionally: $s = x + \epsilon$
- Mean of scores: $\bar{s} = \dfrac{1}{n} \sum_i s_i$

Standard Error of the Mean

- Our scores can come from any distribution; how can we say anything about error bounds if we don’t know this distribution?
- CLT to the rescue!
- CLT says mean of i.i.d random variables with finite mean and variance converges can be approximated with standard normal



Central Limit Theorem: Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \qquad \text{where } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Then the distribution function of U_n converges to the standard normal distribution function as $n \to \infty$. That is,

$$\lim_{n\to\infty} P(U_n \le u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad \text{for all } u.$$

Standard Error of the Mean

- So the estimate of our mean can be transformed into a standard normal
- We can then also get unbiased estimator of sample variance:

$$\mathrm{Var}(s) = \frac{1}{n-1}\sum_i \left(s_i - \bar{s}\right)^2$$

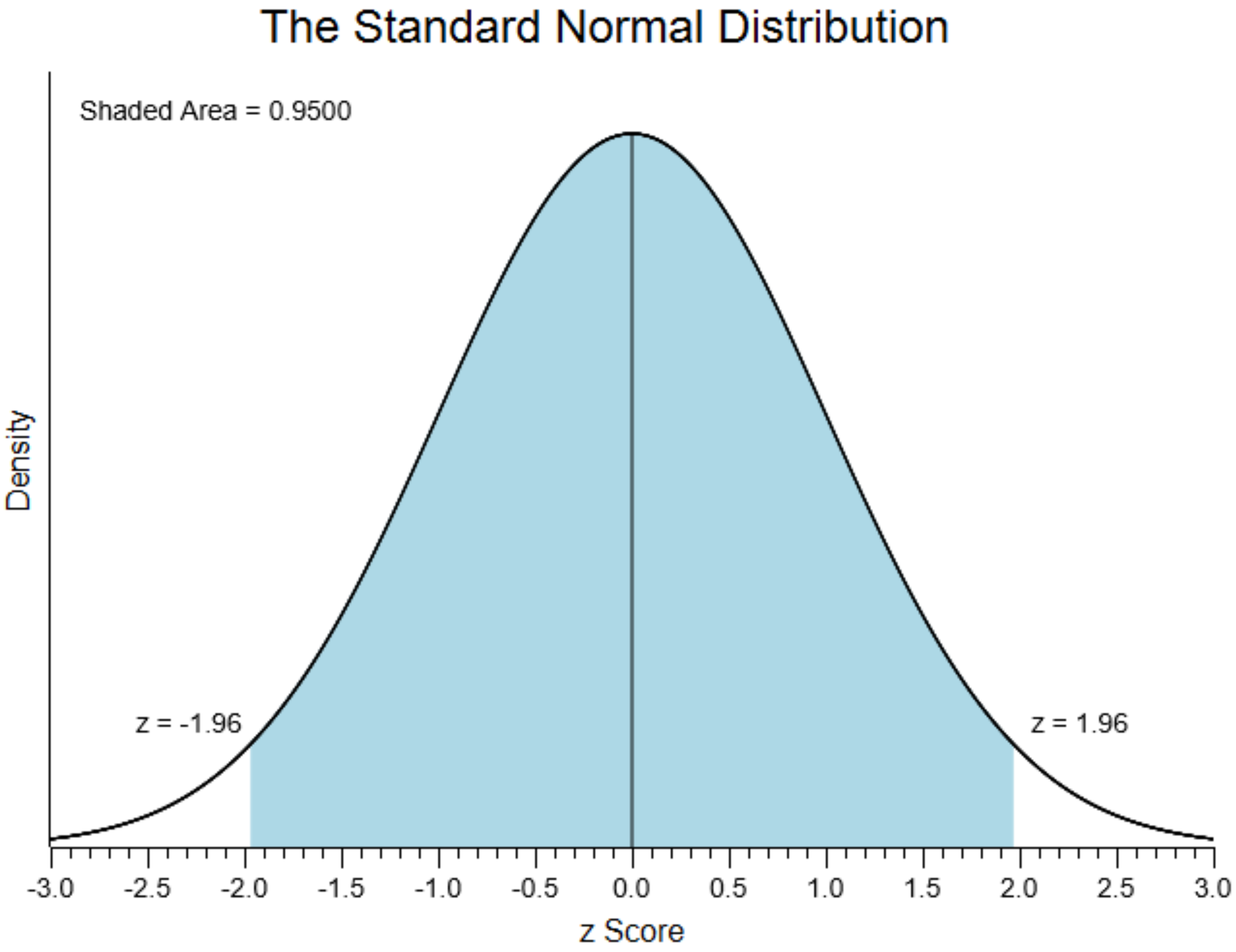
- For n samples, by linearity of variance we recover

$$\mathrm{SE}_{\mathrm{C.L.T.}} = \sqrt{\mathrm{Var}(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_i (s_i - \bar{s})^2\right)/n} \tag{1}$$

Standard Error of the Mean

- Using maximum likelihood estimator (MLE), we declare $\bar{s}$ to be the estimate of population mean, and draw a 95% confidence interval around it (1.96 sigma)
- Recovers Eq (3):

$$\mathrm{CI}\ _{95\%} = \bar{s} \pm 1.96 \times \mathrm{SE}_{\mathrm{C.L.T.}}$$



Clustered Standard Errors

Clustered Standard Error

- CLT requires i.i.d assumption
- Some datasets are clearly not i.i.d
- MGSM (Multilingual Grade-School Math):
  - 2500 grade-school math questions
  - But really: 250 questions translated into 10 different languages
  - 250 clusters of 10

Clustered Standard Error

$$\mathrm{SE}_{\mathrm{C.L.T.}} = \sqrt{\mathrm{Var}(s)/n}$$

- Why does it fail if observations not i.i.d?
  - “Effective” number of observations much fewer than 2500, probably more like 250
- Case 1: observation in each cluster is iid (implies 0 covariance)
- Then

$$\mathrm{SE}_{\text{clustered}} = \sqrt{\mathrm{SE}^2_{\text{C.L.T.}} + \underbrace{\cancel{\frac{1}{n^2}\sum_c \sum_i \sum_{j \neq i} \left(s_{i,c} - \bar{s}\right)\left(s_{j,c} - \bar{s}\right)}}_{=0}}$$

Clustered Standard Error

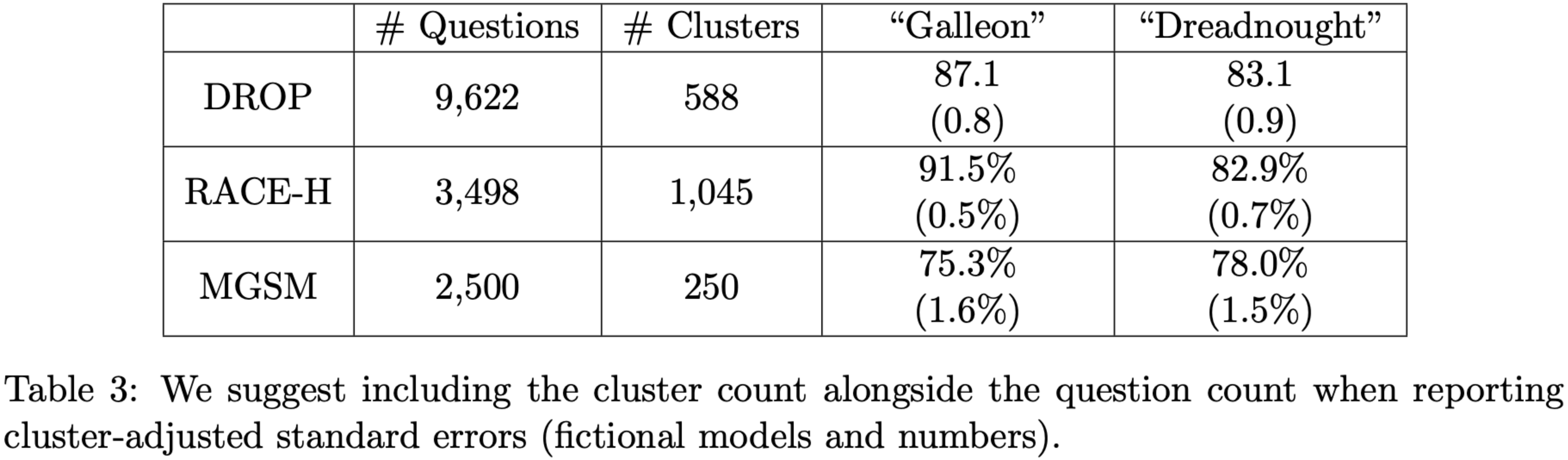
- Case 2: observation in each cluster perfectly correlated
- Then

$$\mathrm{SE}_{\text{clustered}} = \sqrt{\mathrm{SE}^2_{\text{C.L.T.}} + \frac{1}{n^2}\sum_{c}\sum_{i}\sum_{j\neq i}\left(s_{i,c} - \bar{s}\right)^2}$$

and you add back variance contributions within each cluster

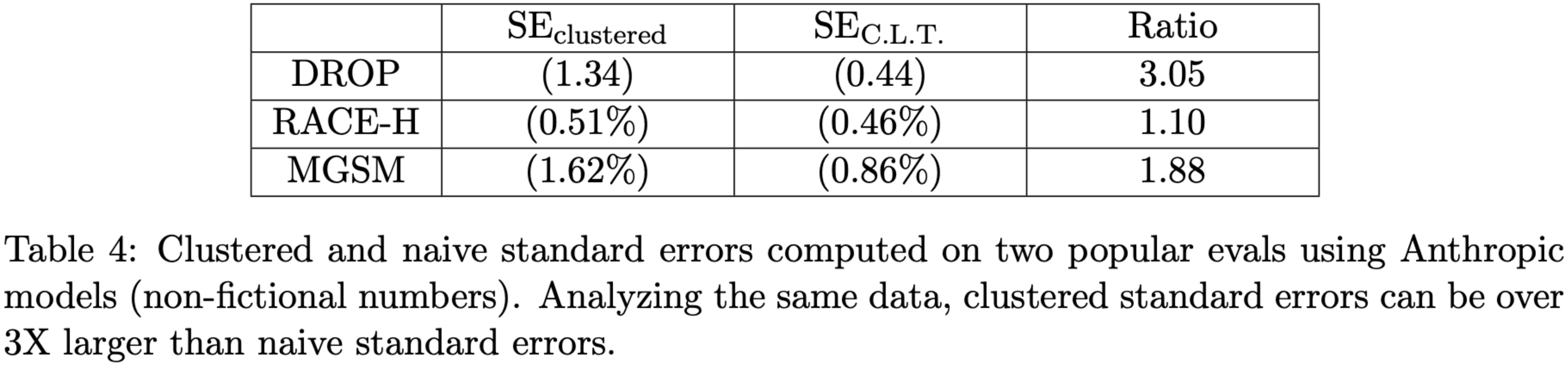
$$\mathrm{SE}_{\text{C.L.T.}} = \sqrt{\mathrm{Var}(s)/n} = \sqrt{\left(\frac{1}{n-1}\sum_{i}(s_i - \bar{s})^2\right)/n}$$

Recommendation for reporting errors



	# Questions	# Clusters	“Galleon”	“Dreadnought”
DROP	9,622	588	87.1 (0.8)	83.1 (0.9)
RACE-H	3,498	1,045	91.5% (0.5%)	82.9% (0.7%)
MGSM	2,500	250	75.3% (1.6%)	78.0% (1.5%)

Table 3: We suggest including the cluster count alongside the question count when reporting cluster-adjusted standard errors (fictional models and numbers).



	$\text{SE}_{\text{clustered}}$	$\text{SE}_{\text{C.L.T.}}$	Ratio
DROP	(1.34)	(0.44)	3.05
RACE-H	(0.51%)	(0.46%)	1.10
MGSM	(1.62%)	(0.86%)	1.88

Table 4: Clustered and naive standard errors computed on two popular evals using Anthropic models (non-fictional numbers). Analyzing the same data, clustered standard errors can be over 3X larger than naive standard errors.

Variance Reduction

Variance Reduction

- $\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left(\frac{1}{n}\sum_i s_i\right) = \mathrm{Var}(s)/n$
- Increase number of samples directly reduces variance
- But we still have another trick..

Law of Total Variance

- This is tricky to get intuition on
- $\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathrm{E}[Y \mid X]$
- Example: Y is dog's weight, X is breed
- First term: avg of variance of weight within each breed (within-group variance)
- Second term: variance of avg of each breed (between-group variance)

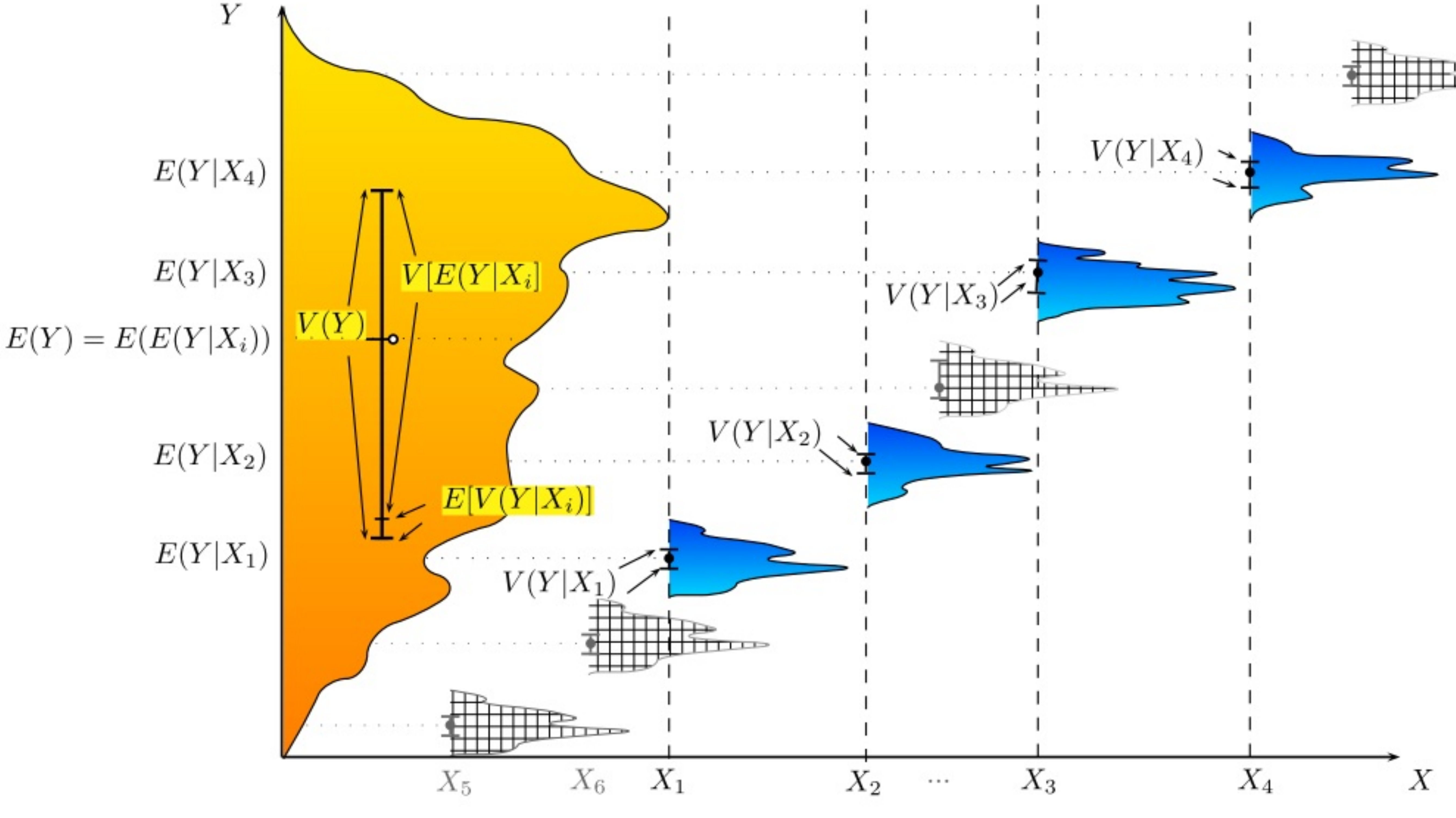
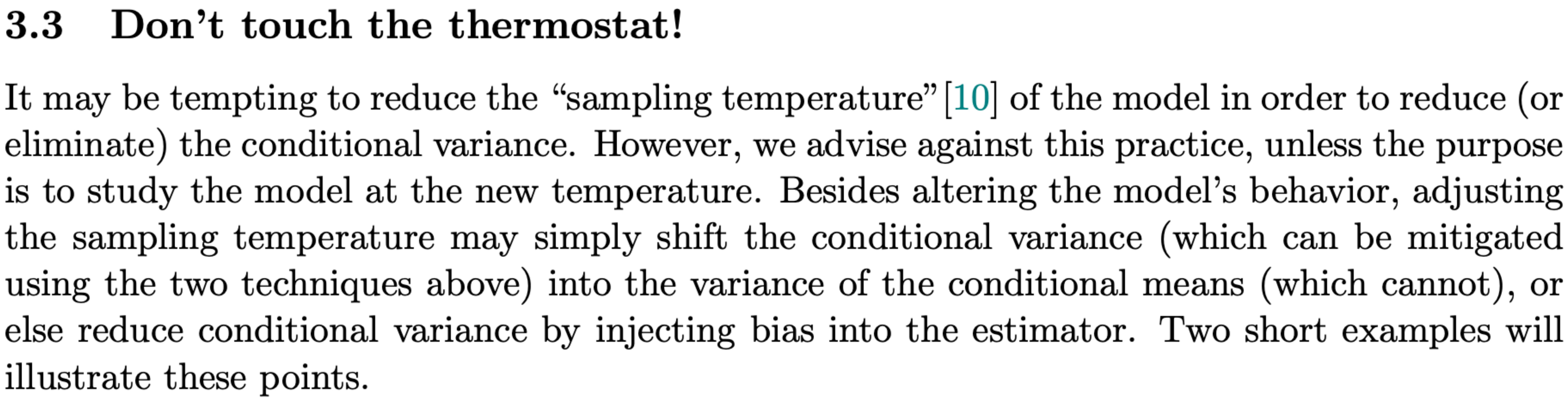


Figure 3: ANOVA : very good fit

Variance Reduction

- FYI: I don't like their notation for this part, very imprecise
- $$\mathrm{Var}(s) = \underbrace{\mathrm{Var}\left(\mathbb{E}[x_i \mid i]\right)}_{\text{variance in scores across different questions}} + \underbrace{\mathbb{E}\left[\mathrm{Var}(x_i \mid i)\right]}_{\text{variance in scores from answering the same question across different attempts}}$$
- Let's consider resampling
- Resampling won't help the first term - this is inherent in the distribution of questions
- But it can help to decrease the second term: sampling n times & taking mean will reduce it by n
- Increasing n is economical until the point that second term is same size as first term (then first term dominates)

Variance Reduction

- $$\mathrm{Var}(s) = \underbrace{\mathrm{Var}\left(\mathbb{E}[x_i \mid i]\right)}_{\text{variance in scores across different questions}} + \underbrace{\mathbb{E}\left[\mathrm{Var}(x_i \mid i)\right]}_{\text{variance in scores from answering the same question across different attempts}}$$
- Tempting thing to eliminate second term: set temp=0
- 

  3.3 Don’t touch the thermostat!

  It may be tempting to reduce the “sampling temperature” [10] of the model in order to reduce (or eliminate) the conditional variance. However, we advise against this practice, unless the purpose is to study the model at the new temperature. Besides altering the model’s behavior, adjusting the sampling temperature may simply shift the conditional variance (which can be mitigated using the two techniques above) into the variance of the conditional means (which cannot), or else reduce conditional variance by injecting bias into the estimator. Two short examples will illustrate these points.
- Their example: setting T=0 increases first term

Variance Reduction

- $\mathrm{Var}(s) = \underbrace{\mathrm{Var}\left(\mathbb{E}[x_i \mid i]\right)}_{\text{variance in scores across different questions}} + \underbrace{\mathbb{E}\left[\mathrm{Var}(x_i \mid i)\right]}_{\text{variance in scores from answering the same question across different attempts}}$
- For problems where you can use model logprobs to get probability of correct answer (i.e true/false qn), here’s another trick:
- Instead of sampling the answer token & giving a binary score, return the probability of correct answer as score
- Then second term becomes 0 😊

Paired Analysis

Comparing Models

- Suppose you are deciding whether model B is better than model A
- You can come up with a new metric that is the difference of their means: $\hat{\mu}_{A-B} = \hat{\mu}_A - \hat{\mu}_B$
- If this difference is positive and large, we should use model B!
- Assuming independence, standard error is $\mathrm{SE}_{A-B} = \sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2}$

Comparing Models

- But this is an instance where non-independence can actually help us: since we evaluate both models on the same test cases in the eval, it is likely both models may find the same groups of test cases similarly easy or challenging
- What they call “paired” is really just whether we are assuming independence or not
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$, and $\mathrm{Cov}(X,Y) = 0$ if they are independent
- Similarly you can expand & work out that $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y)$
- Get variance reduction if there is correlation between the scores of the two models!

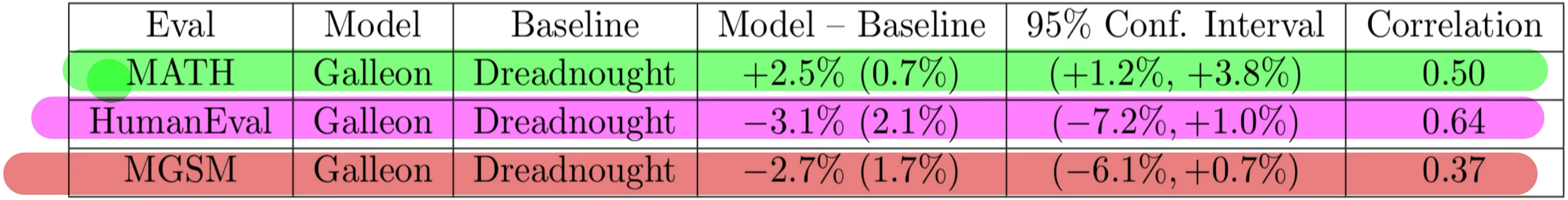
Paired Analysis

Example

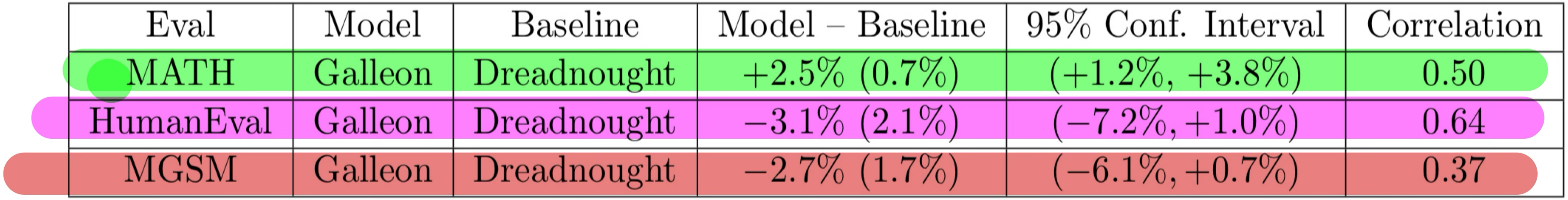
- $\mathrm{Var}\left(\hat{\mu}_{A-B,\text{ unpaired}}\right) = \left(\mathrm{Var}\left(s_A\right) + \mathrm{Var}\left(s_B\right)\right)/n$
- $\mathrm{Var}\left(\hat{\mu}_{A-B,\text{ paired}}\right) = \left(\mathrm{Var}\left(s_A\right) + \mathrm{Var}\left(s_B\right) - 2\,\mathrm{Cov}\left(s_A, s_B\right)\right)/n$
- If $\mathrm{Var}(s_A) = \mathrm{Var}(s_B) = 1/12$ and $\mathrm{Cov}(s_A, s_B) = 1/24$, then unpaired analysis has 1/6n variance whereas paired analysis gives 1/12 variance - a 50% reduction!

Paired Analysis

- Recommendation:
  - Pairwise difference & standard error
  - Score correlations

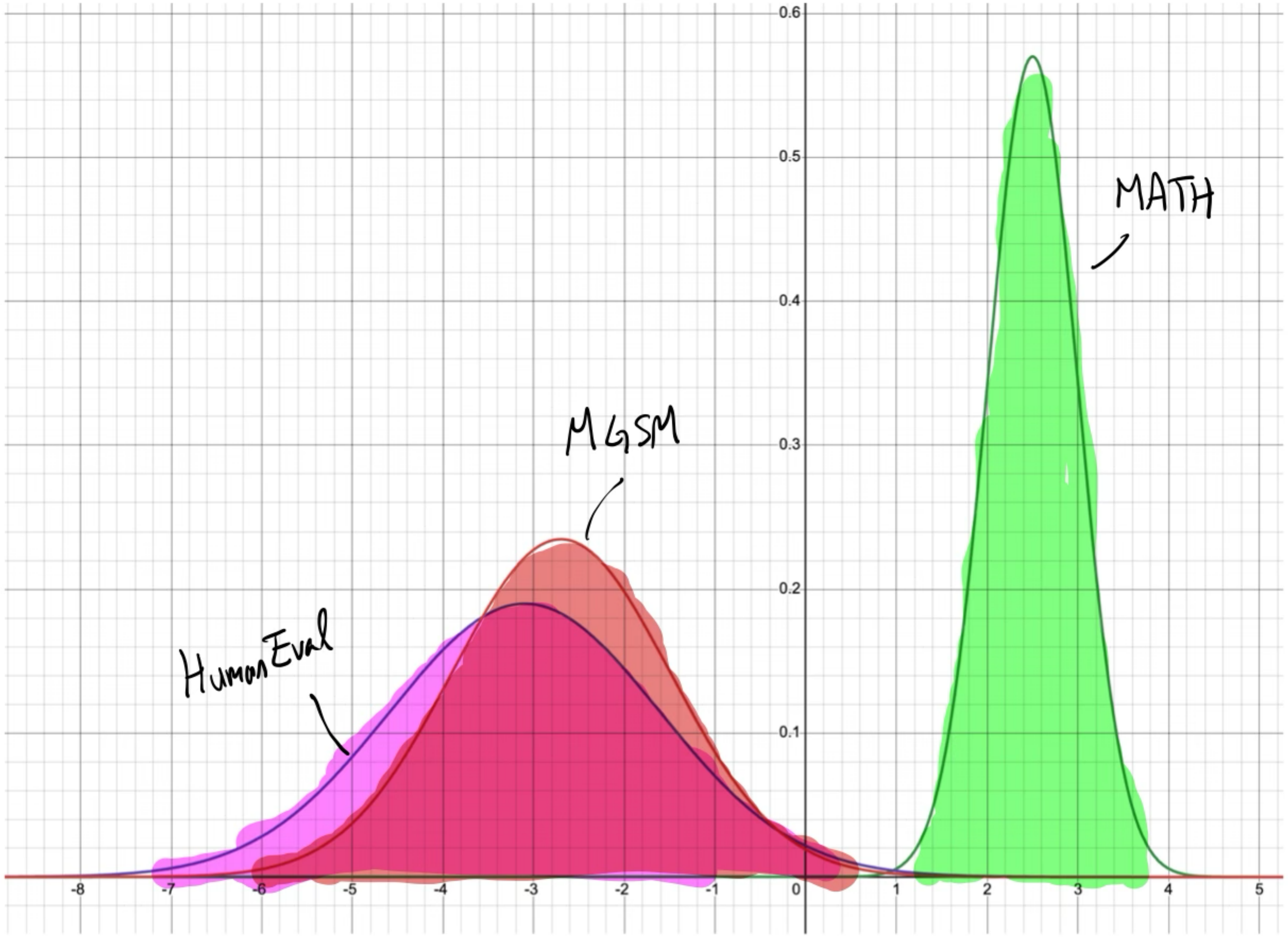


Eval	Model	Baseline	Model – Baseline	95% Conf. Interval	Correlation
MATH	Galleon	Dreadnought	+2.5% (0.7%)	(+1.2%, +3.8%)	0.50
HumanEval	Galleon	Dreadnought	−3.1% (2.1%)	(−7.2%, +1.0%)	0.64
MGSM	Galleon	Dreadnought	−2.7% (1.7%)	(−6.1%, +0.7%)	0.37



Eval	Model	Baseline	Model – Baseline	95% Conf. Interval	Correlation
MATH	Galleon	Dreadnought	+2.5% (0.7%)	(+1.2%, +3.8%)	0.50
HumanEval	Galleon	Dreadnought	−3.1% (2.1%)	(−7.2%, +1.0%)	0.64
MGSM	Galleon	Dreadnought	−2.7% (1.7%)	(−6.1%, +0.7%)	0.37

- Model beats baseline on MATH (95% confidence interval of difference all in positive region)
- The other two…nope



Power Analysis

Crash Course in Statistical Testing

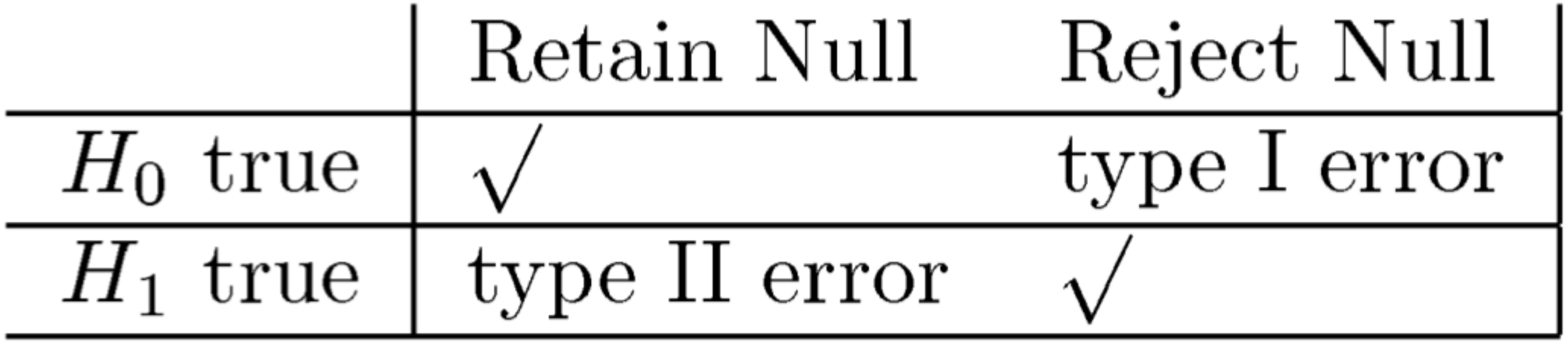
- Suppose Jones claims he will get more than 50% of the votes in the city election (null hypothesis)
- We don’t believe this, and want to show the hypothesis that Jones has <50% of the votes (alternative hypothesis)
- To do so, we try to show the null hypothesis is unlikely based on data.
  - If we can do this, we can reject the null hypothesis and conclude the alternative is probably true
  - If we can’t, we do not accept the alternative - we reserve judgement and state that there is insufficient evidence to conclude that the alternative is probably true.
- Note that you never try to prove something in statistics, you can only reject hypothesis based on data

Crash Course in Statistical Testing

- Elements of a statistical test
  - Null hypothesis H_0: $p = 0.5$
  - Alternative hypothesis H_a or H_1: $p < 0.5$
  - Test statistic: Y, the number of people who voted for Jones from a sample of 15 people
  - Rejection region: $\{Y \leq k\}$ for some threshold k

Quick aside: Types of errors

- Type I error (α): false positive. Also called the significance level of the test
  - Diagnosing a healthy person as diseased
  - Convicting an innocent person
- Type II error (β): false negative
  - Failing to diagnose a diseased person as sick
  - Acquiring a guilty person



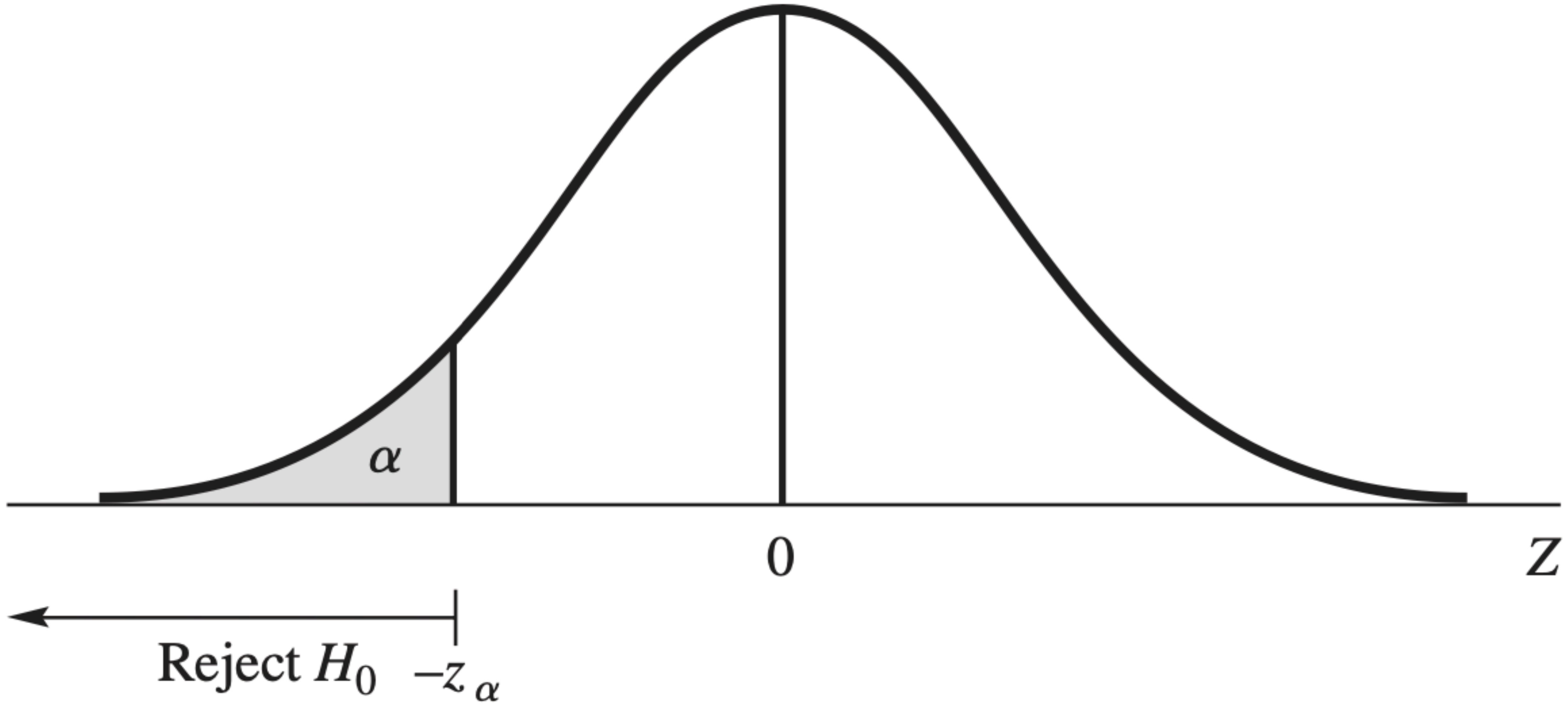
	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

Crash Course in Statistical Testing

- Type I error (α): rejecting H_0 when it is actually true
- Type II error (β): accepting H_0 when H_a is true
- Interesting question: how to choose k for 15 voters?
- Choose small $k = 2$ (only claim he will lose if he gets 2 votes or less)
  - Low chance of committing Type I error (it's likely he will lose when we think so)
  - High chance of committing Type II error (he will also frequently lose when we don't think so)
- Choose large $k = 5$ (claim he will lose even if he has up to 5 votes)
  - High chance of committing Type I error (he will win many times we think he will lose)
  - Low chance of committing Type II error (if we think he will win, he'll likely win)
- α and β are inversely related!

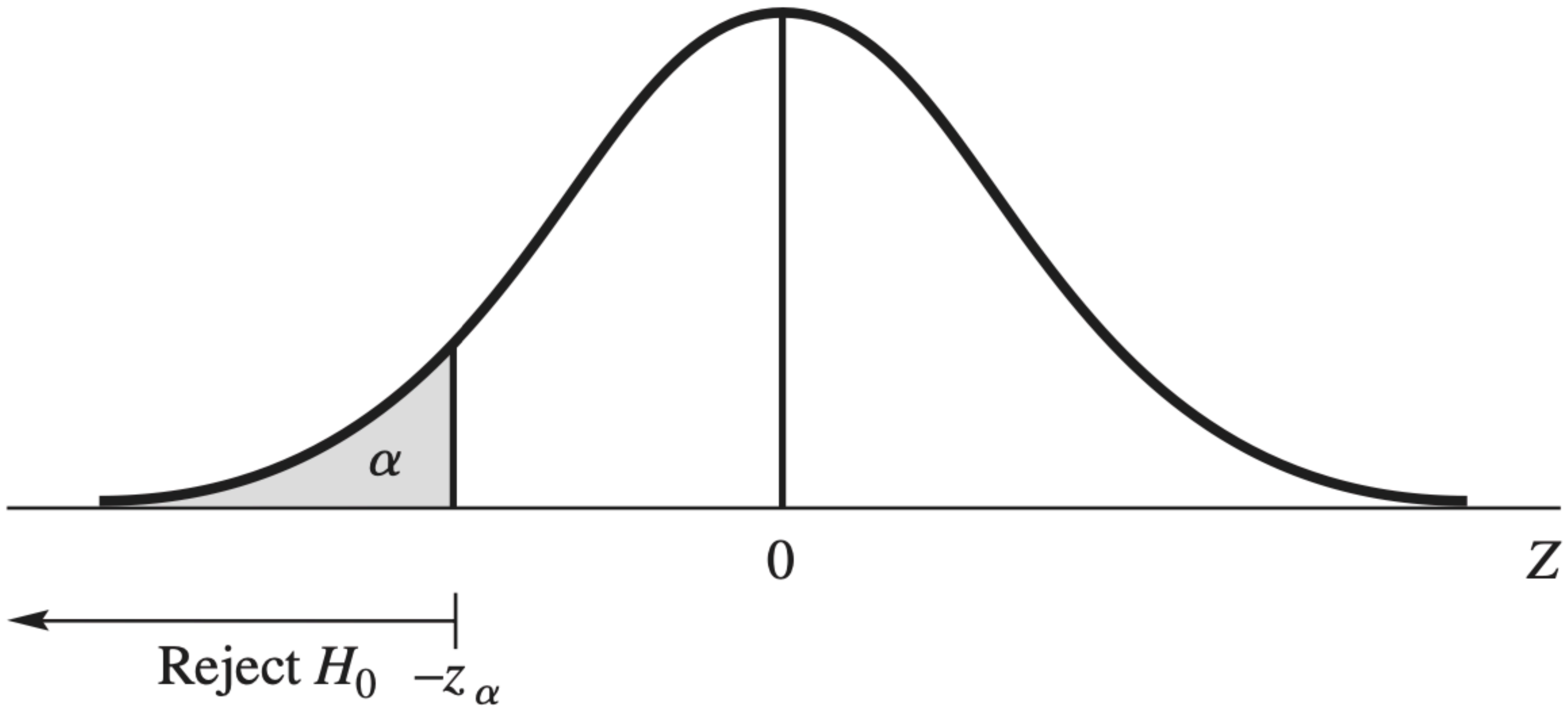
Choosing k

- How to choose rejection region k in practice:
- Choose a α (i.e 0.05)
- Compute the k that gives desired rejection region with area α
- Can also do a similar thing if you want to control for β instead (though much harder)
    - But we usually care more about Type I errors, i.e wrongly claiming something works when it actually does not in science



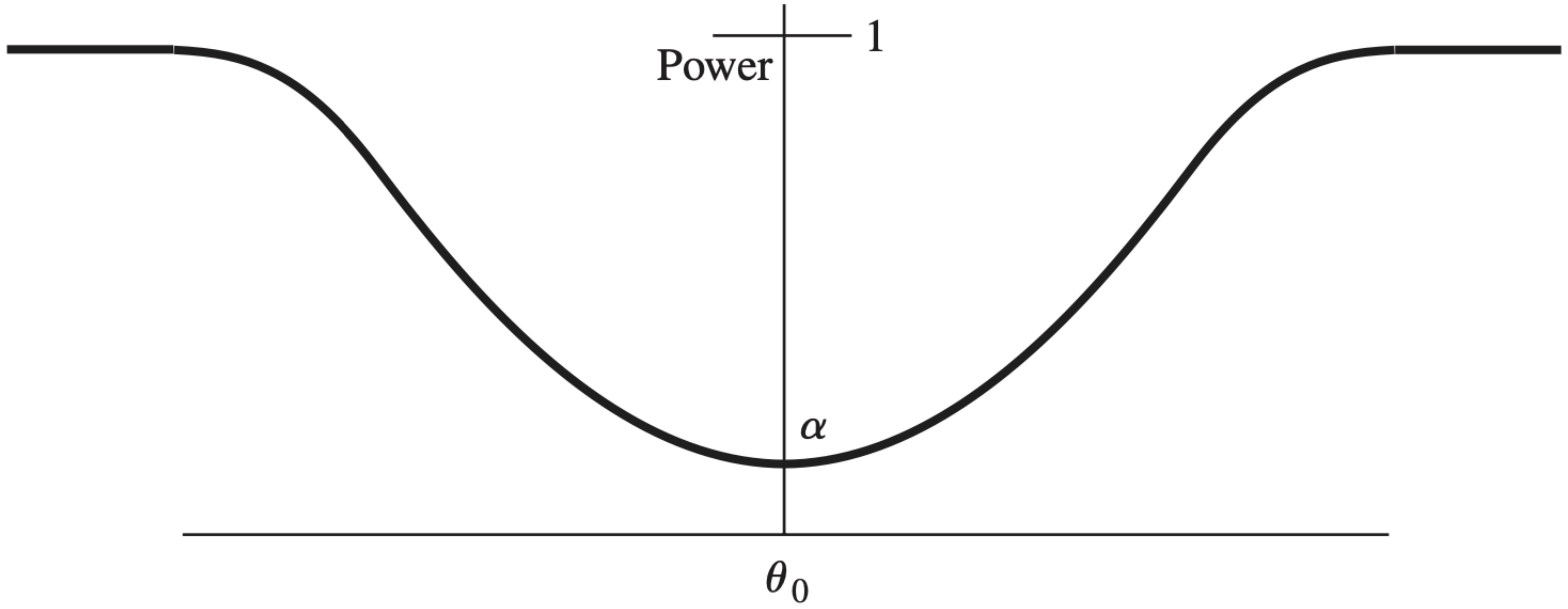
Choosing k

- Of course, you could have a weird experiments where test statistic is in rejection region at α = 0.05 but not so at α = 0.0499, which makes this choice somewhat arbitrary
- So people also report p-value: the smallest α such that the test statistic is still in the rejection region



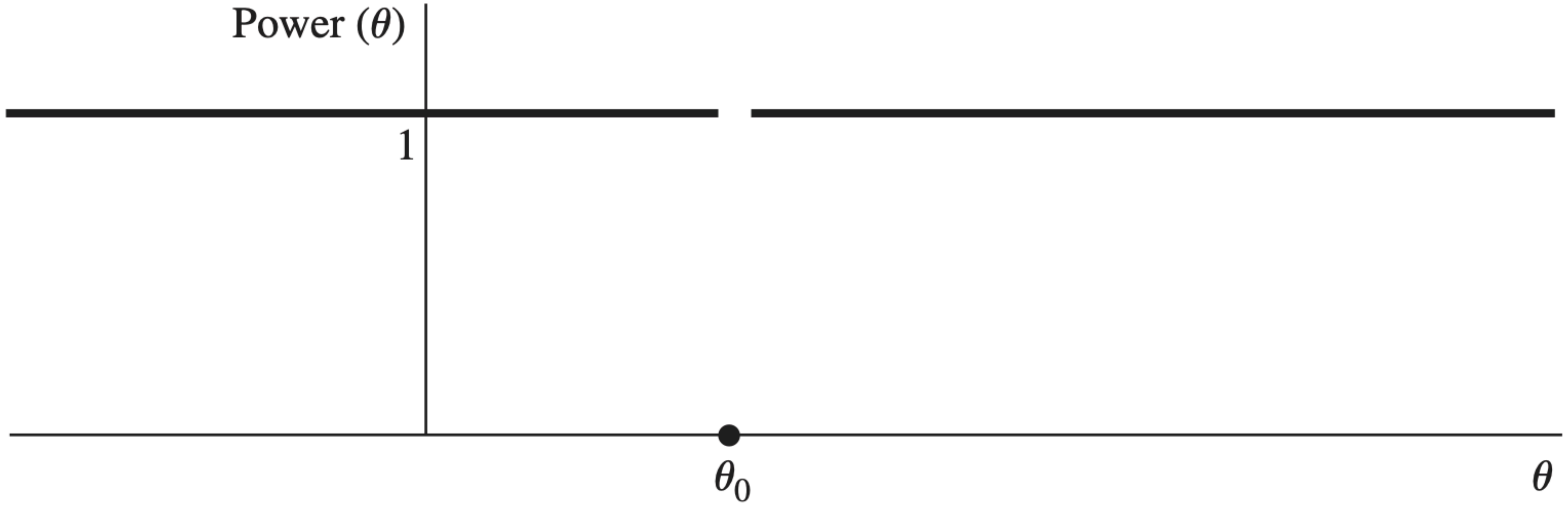
Power Analysis

- Goodness of a test is measured by α and β
- Power of a test is the probability it will lead to rejection of H_0
- $\text{power}(\theta) = P(W$ in rejection region when parameter value is $\theta)$
- Typical power curve:
  - Low probability of rejection when true parameter close to θ_0, and vice versa



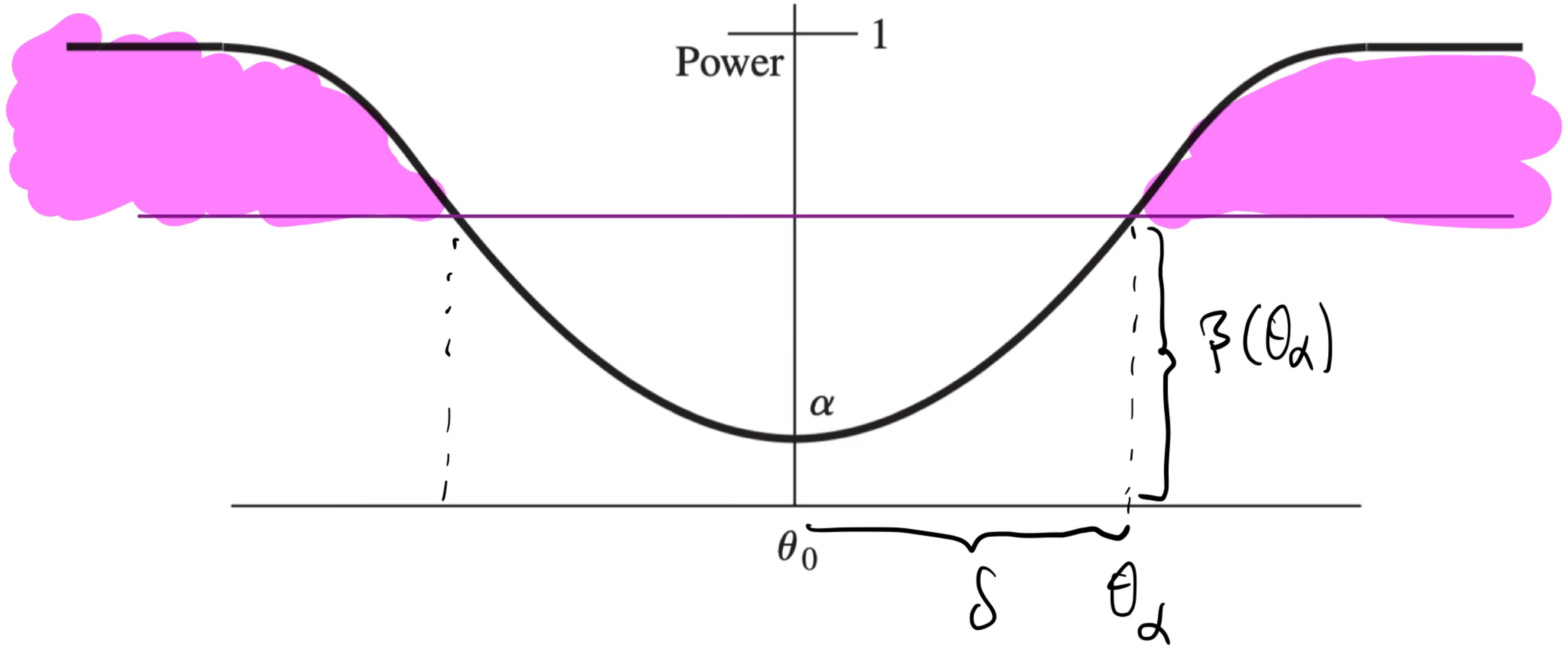
Power Analysis

- In an ideal world
  - Never reject H_0 if true parameter is θ_0
  - Always reject otherwise



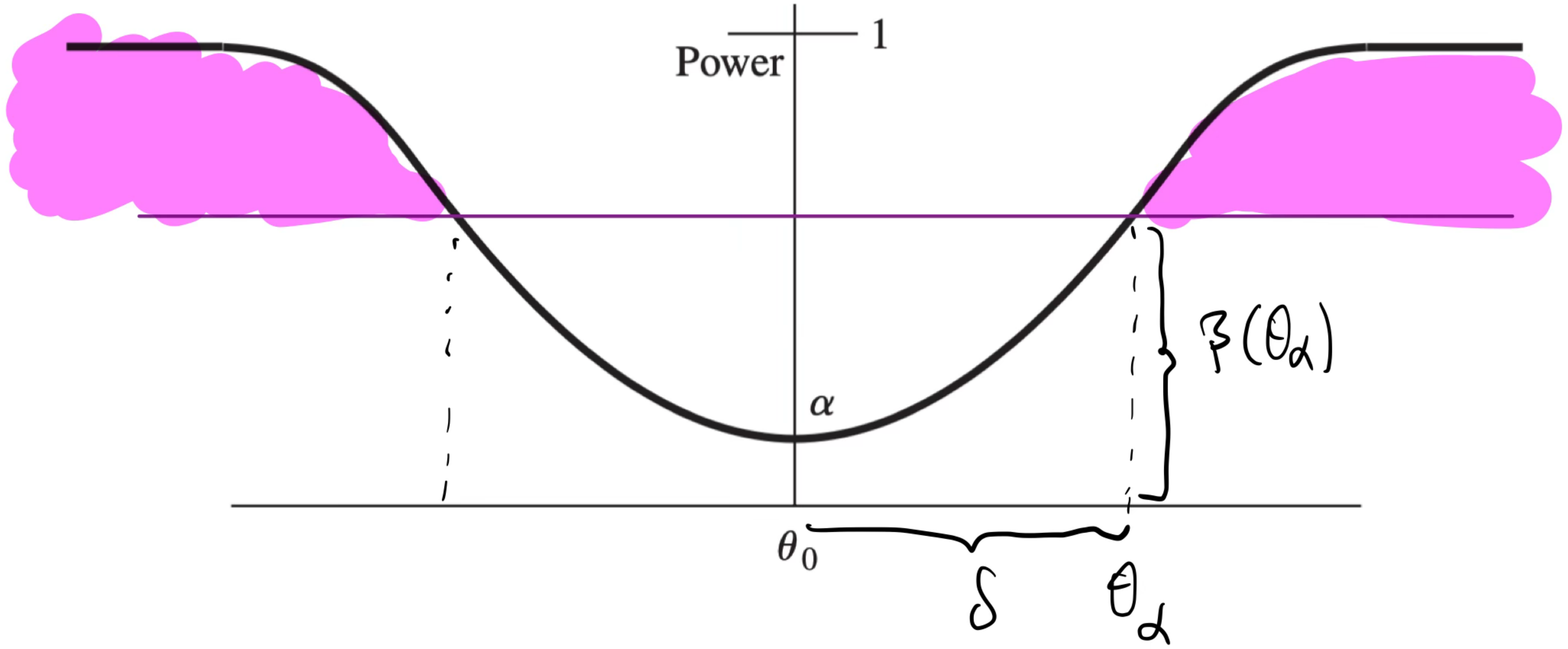
Power Analysis

- So tests with higher power would be able to detect small changes
- Conversely, if your test has low power at H_A, then it's pointless to spend money running evals like "is model B actually 1% better than model A" because it will rarely reject the null hypothesis when it should
  - Observe that $\beta(\theta_A) = 1 - \text{power}(\theta_A)$
- One can ask: what is the minimum detectable effect (MDE) δ for a desired power level?



Power Analysis

- In essence, Section 5 shows that increasing size of dataset will decrease variance which increases power of test
- So for a given minimum detectable effect, significance level α, and desired power level, you can compute the minimum dataset size needed to make this happen



Further Reading

- Mathematical Statistics with Applications (Wackerly, 7 ed.)
    - Ch 8 Estimation, Ch 10 Hypothesis Testing, Ch 13 The Analysis of Variance
- All of Statistics: A Concise Course in Statistical Inference (Wasserman)
    - Ch 6 Models, Statistical Inference and Learning, Ch 8 The Bootstrap, Ch 10 Hypothesis Testing and p-values