

PSYC 193R:

Data Analysis & Visualization with R

Lecture 6: One-sample t-test (preparation)

Jonas S.-H. Lau [silau -et- ucsd.edu]
psyc193r.ucsd.edu



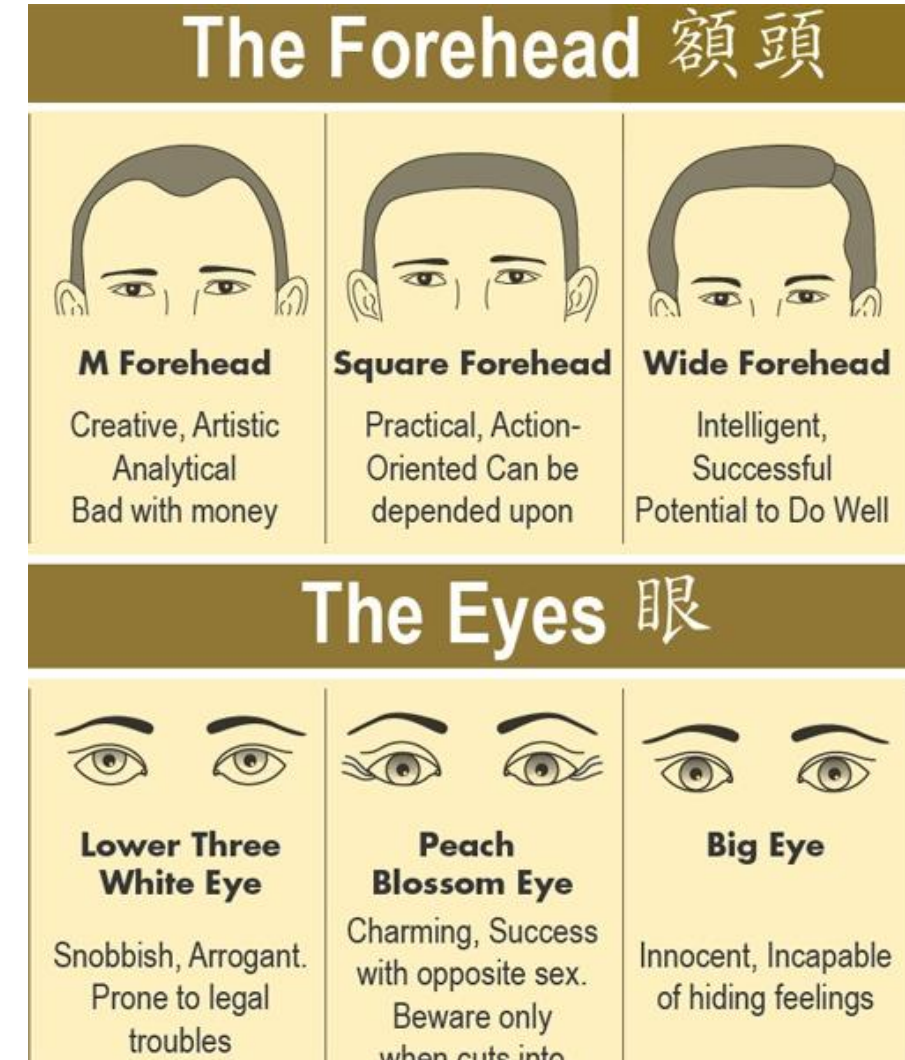
The Art of Face Reading

Jonas wants to know if the traditional art of face reading has a scientific basis.

Needless to say, there will be enormous benefits if we can read people's preferences reliably.

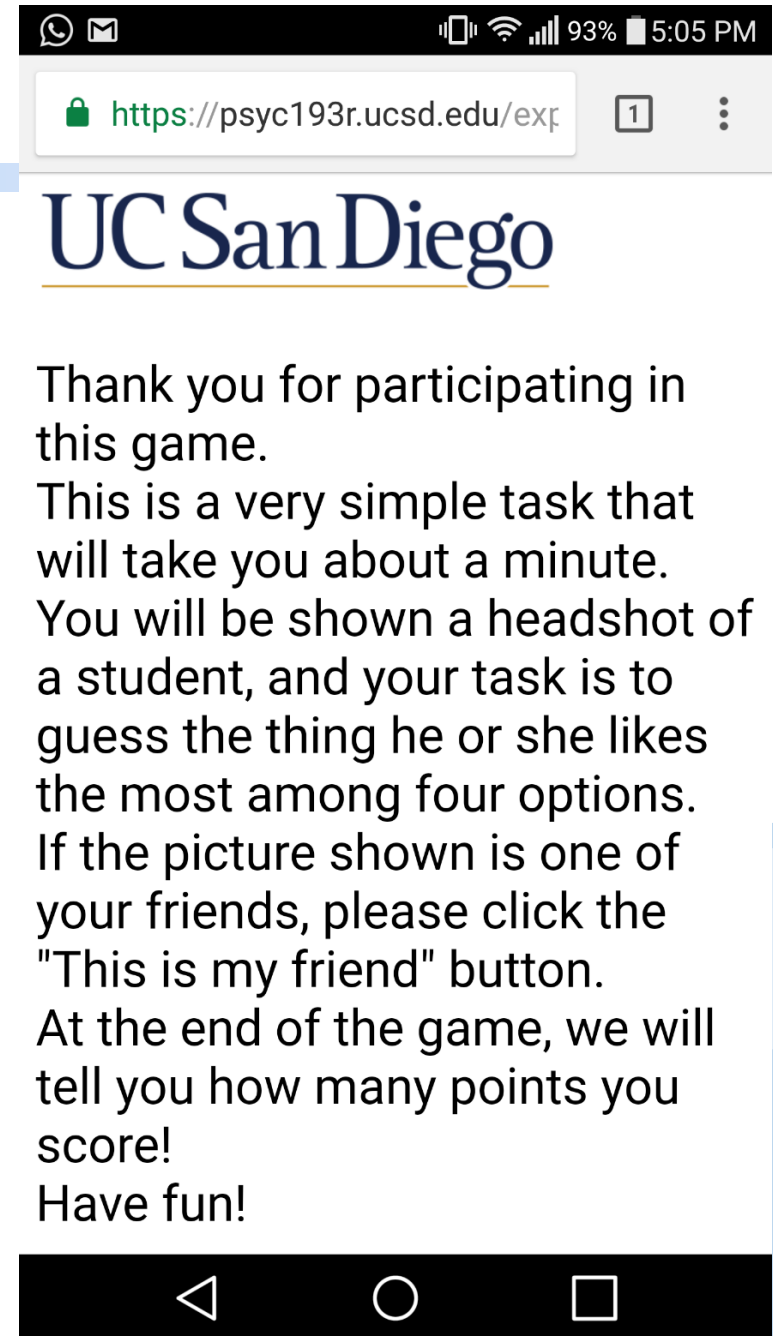
(e.g., first date, job interview, business meetings, ...)

Design a study to scientifically examine the issue.



Face Reading Experiment

- n participants performed 10 – 13 face-reading trials.
- In each trial, a headshot was shown together with four choices. Participants had to guess which the person's preference was out of the four options.
 - Question order was randomized
 - Question choices were randomized



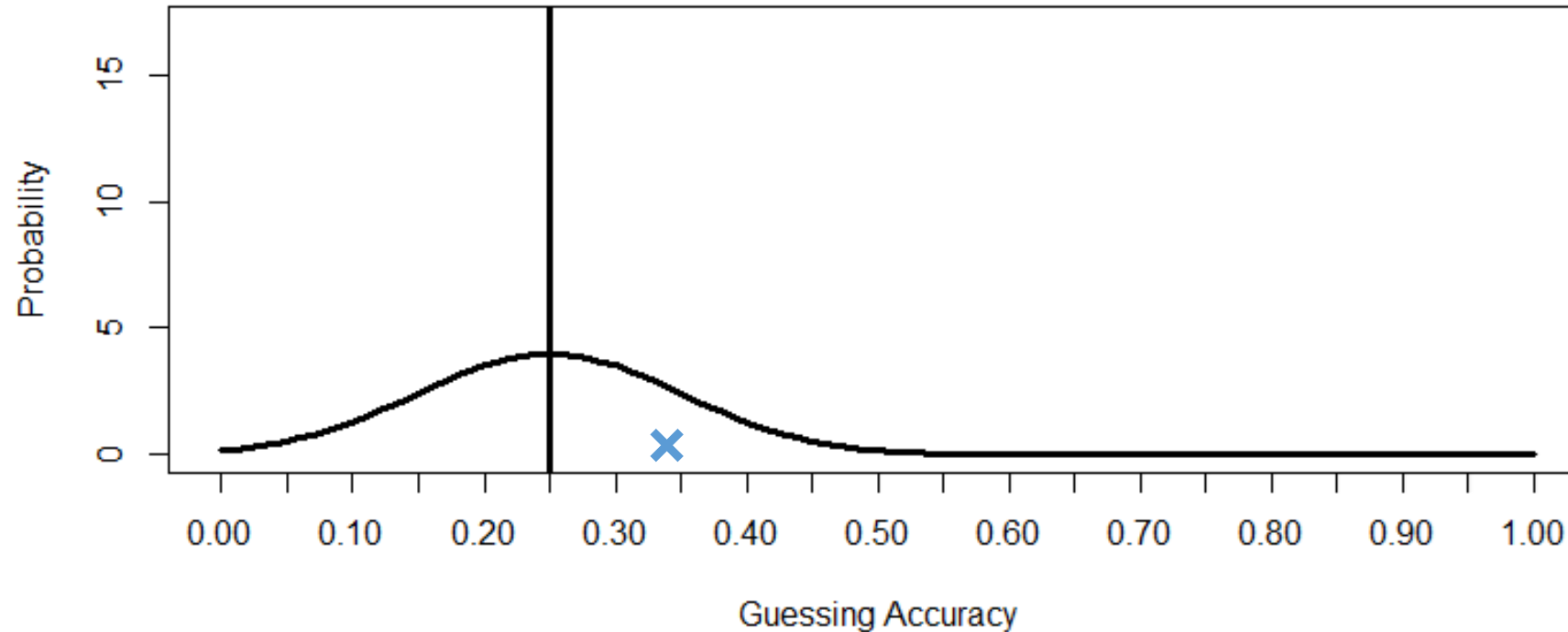
The Design

- What was the research question?
- What was the target population?
- What were the IV(s) and DV(s)?
- How were the IV(s) and DV(s) operationalized?
(also, scale of measurement?)
- Was it a between / within-subjects design?
- What are the expected results?
If the manipulation does not work:
If the manipulation works:



The basic structure

Baseline population distribution (no face / random guessing)



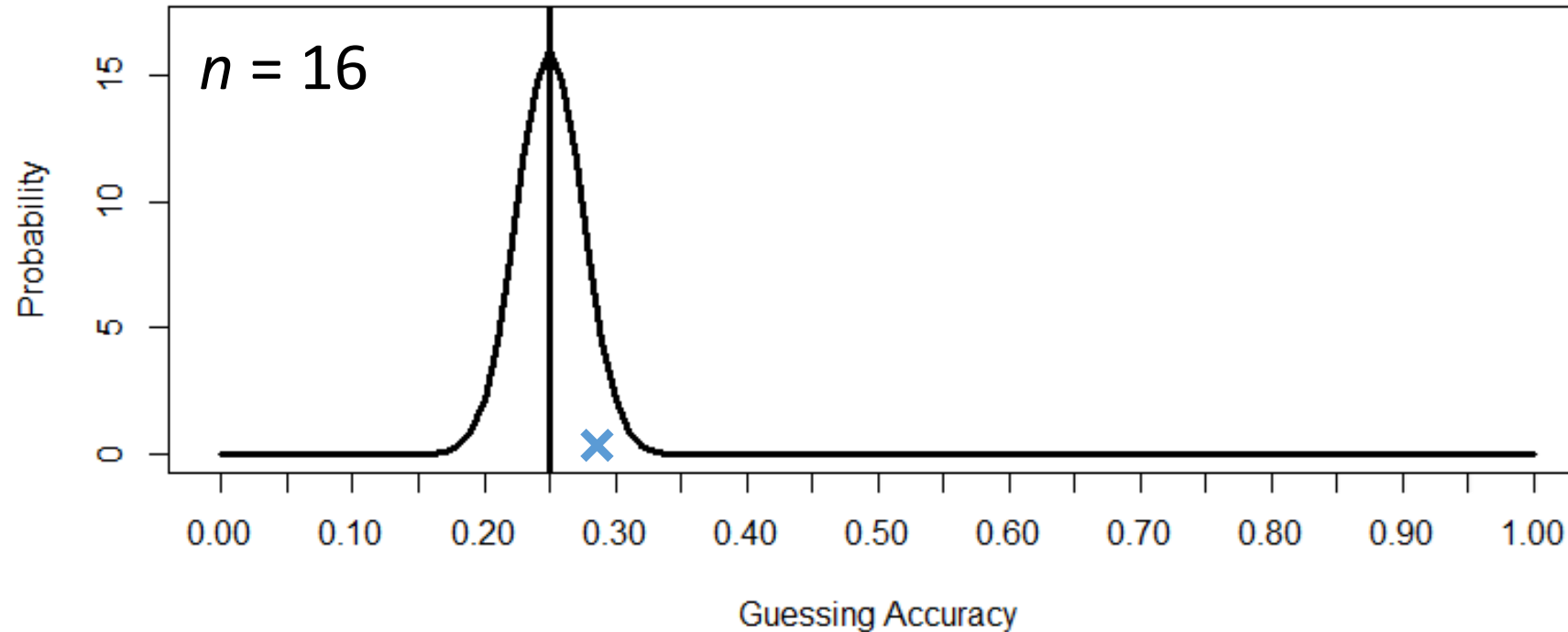
Sample Representing
the treated population
(guess with face)

$n = 16$



The basic structure

Baseline sampling distribution (no face / random guessing)



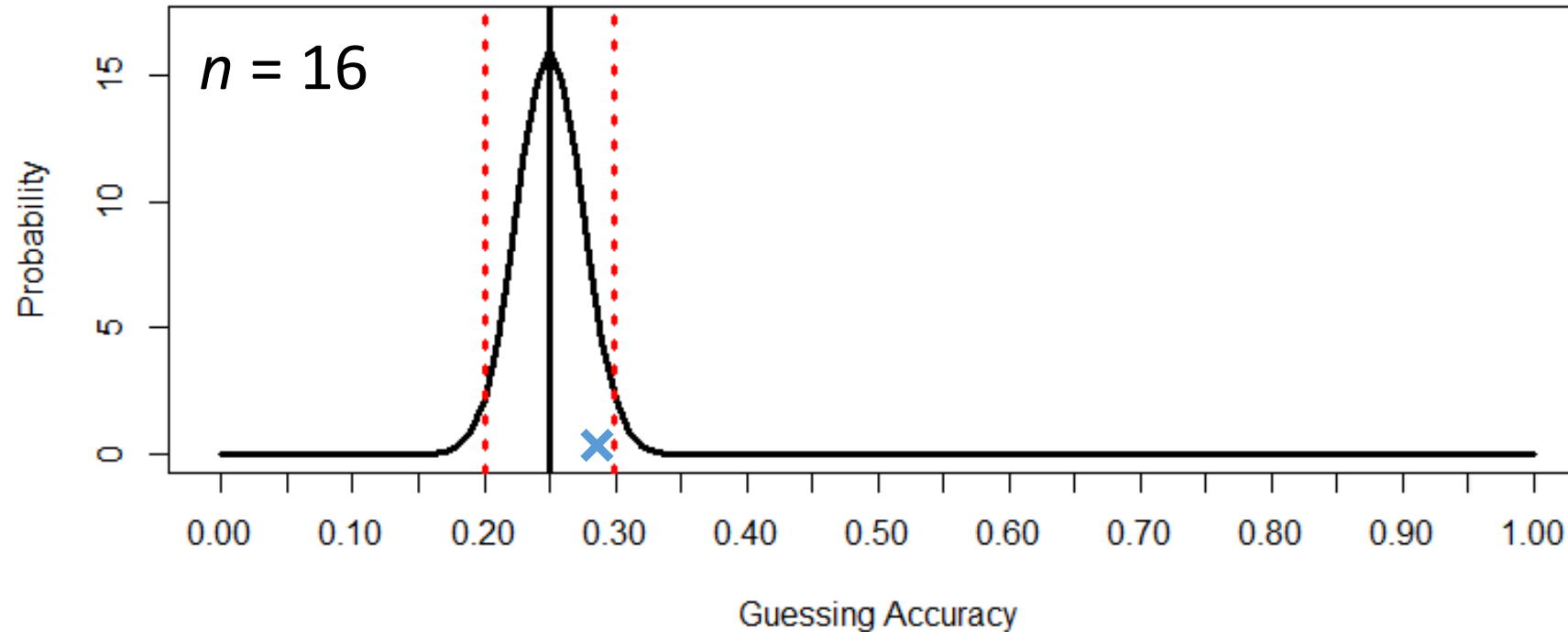
Sample Representing
the treated population
(guess with face)

$n = 16$



The basic structure

Baseline sampling distribution (no face / random guessing)



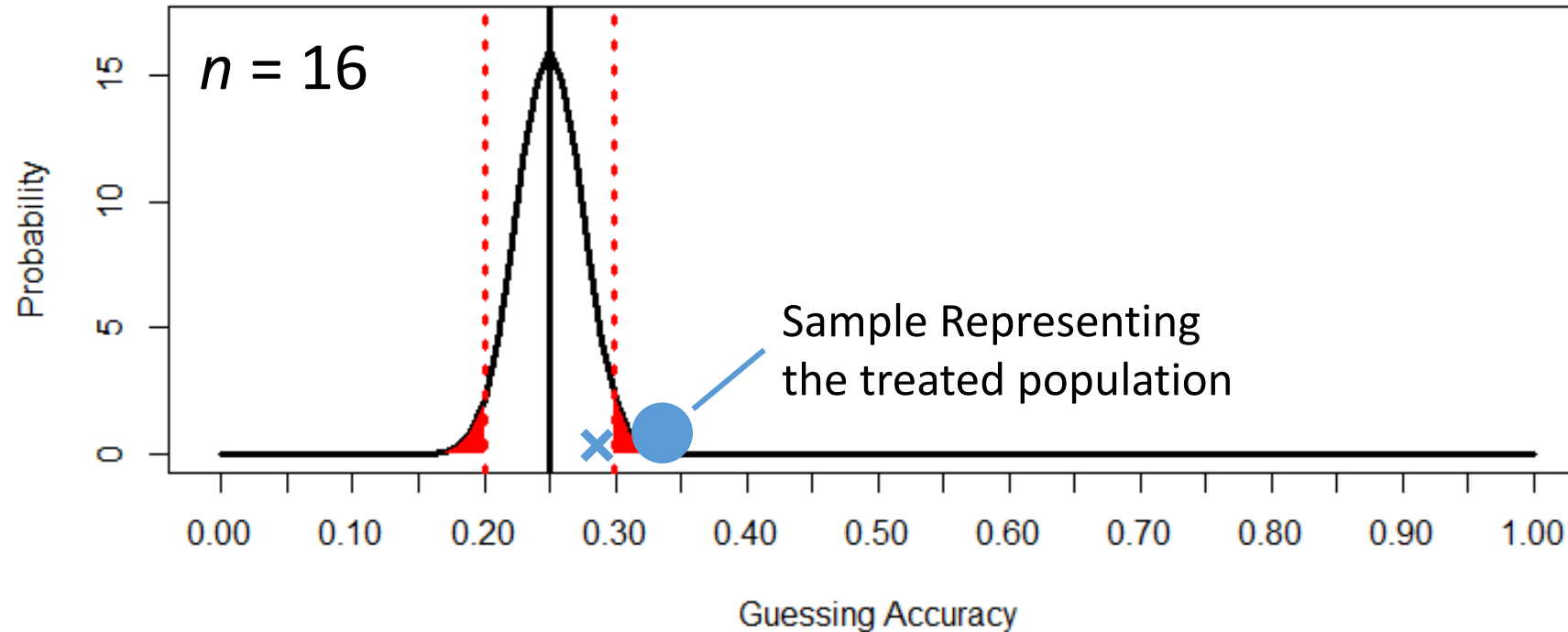
Sample Representing
the treated population
(guess with face)

$n = 16$



The basic structure

Baseline sampling distribution (no face / random guessing)



Z-test

Information needed

- From the baseline population:

$\mu_{baseline}$

σ

kurtosis: normality assumption

- From the sample:

Sample mean $M_{treatment}$

Sample Size n

Actual Information we have

- From the baseline population:

$\mu_{baseline}$ (theoretical)

- From the sample:

From the individual scores...

Sample mean $M_{treatment}$

Sample Size n



Z-test

Information needed

- From the baseline population:

μ_{baseline}

σ

kurtosis

- From the sample:

Sample mean M_{treatm}

Sample Size n

$n \geq 30$

Sampling
distribution
approaches normal

Z distribution

Actual Information we have

- From the baseline population:

μ_{baseline} (theoretical)

- From the sample:

From the individual scores...

Sample mean $M_{\text{treatment}}$

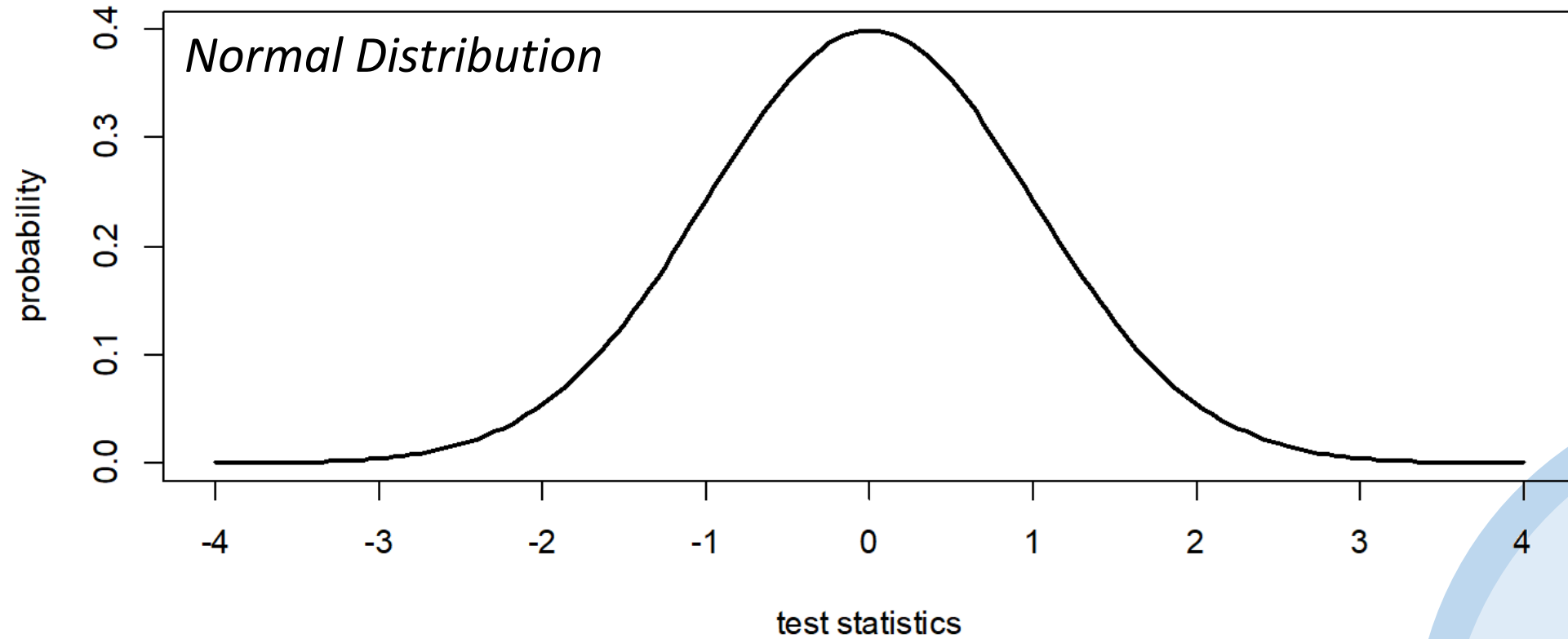
Sample Size n

degrees of freedom $(n - 1)$

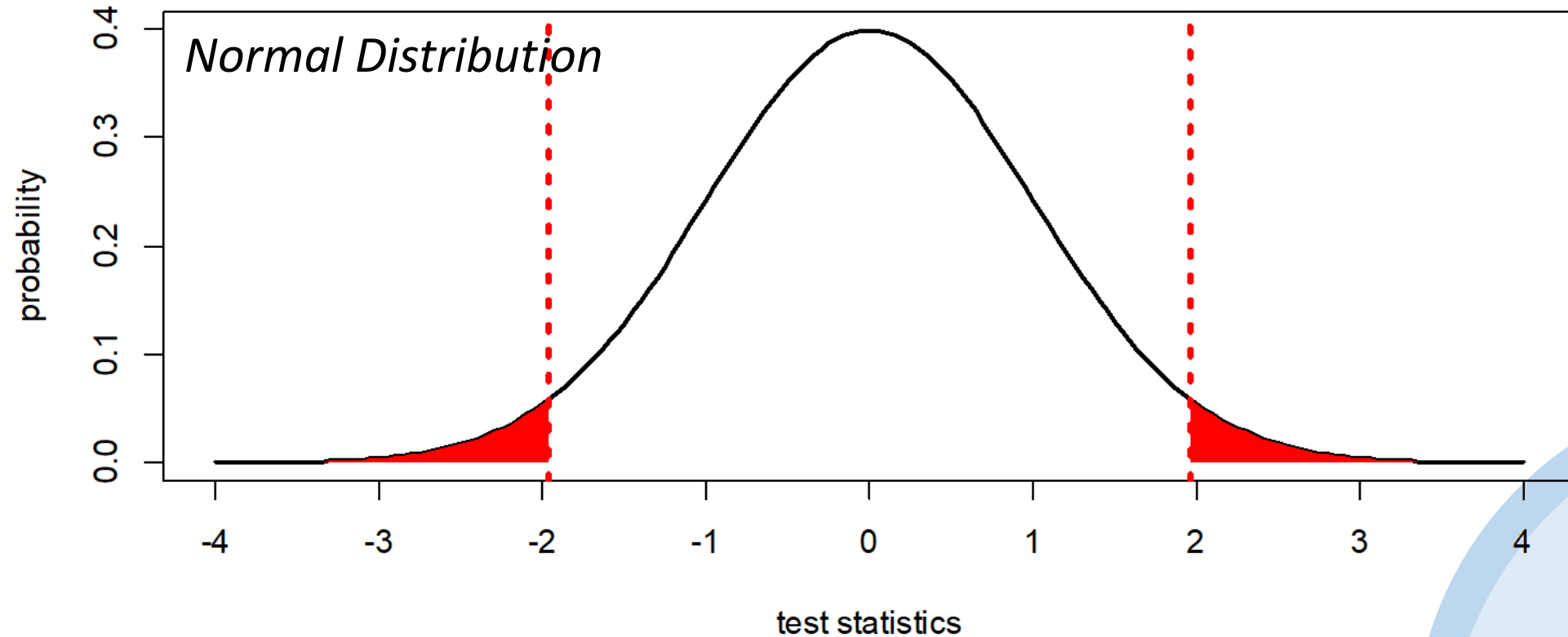
t distribution

$$S = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$$

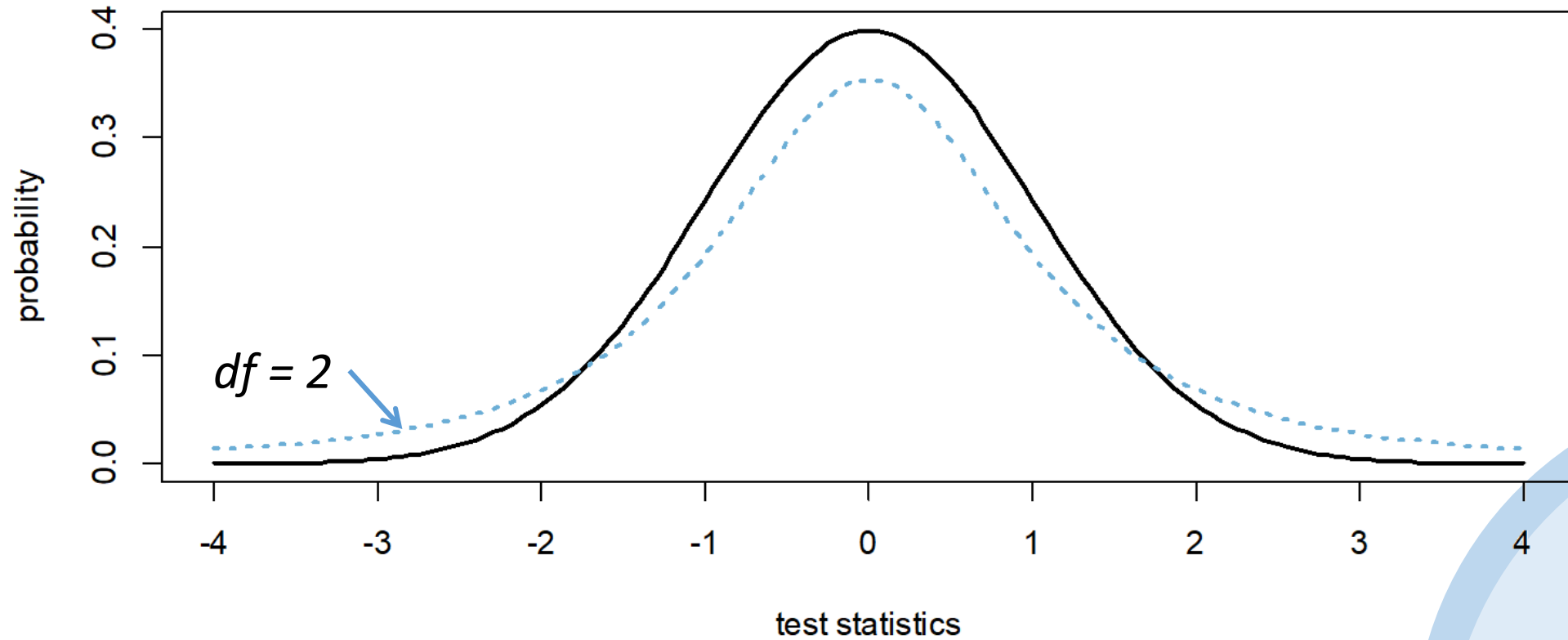
Normal (Z) distribution



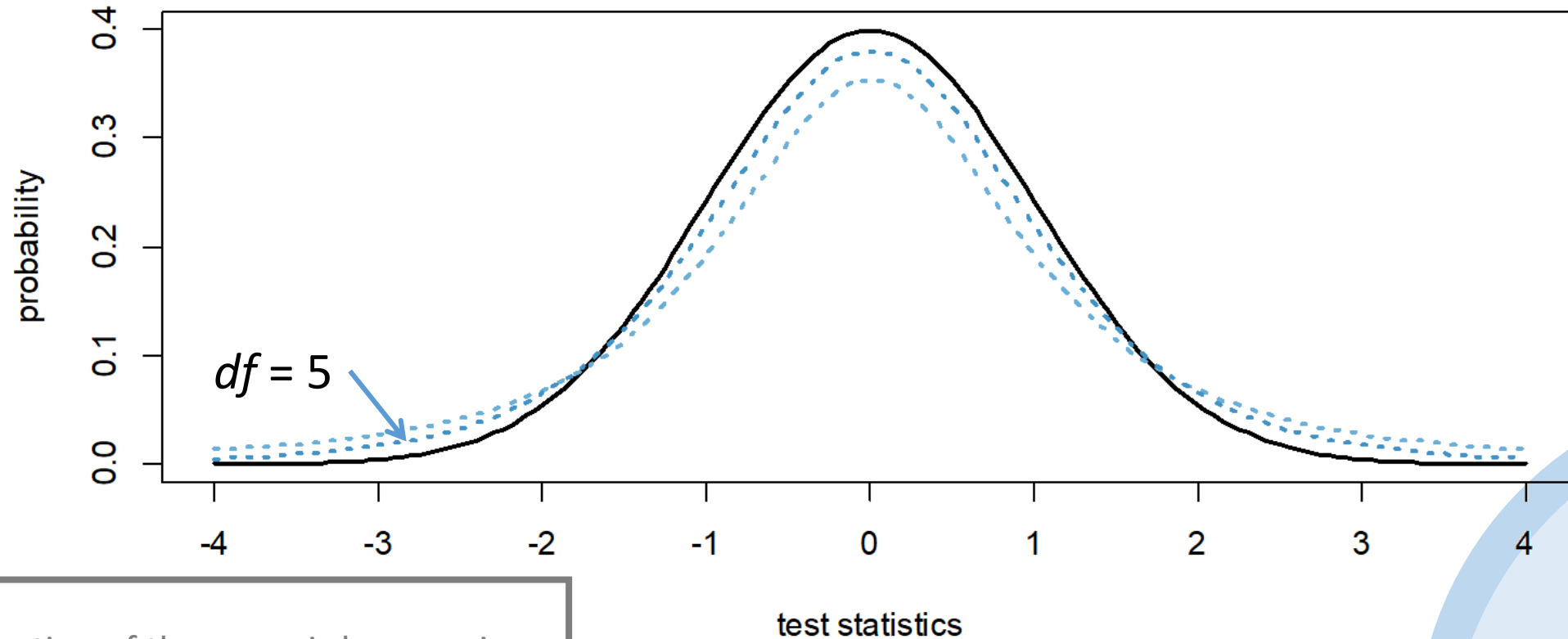
Normal (Z) distribution



t distribution ($df = 2$)

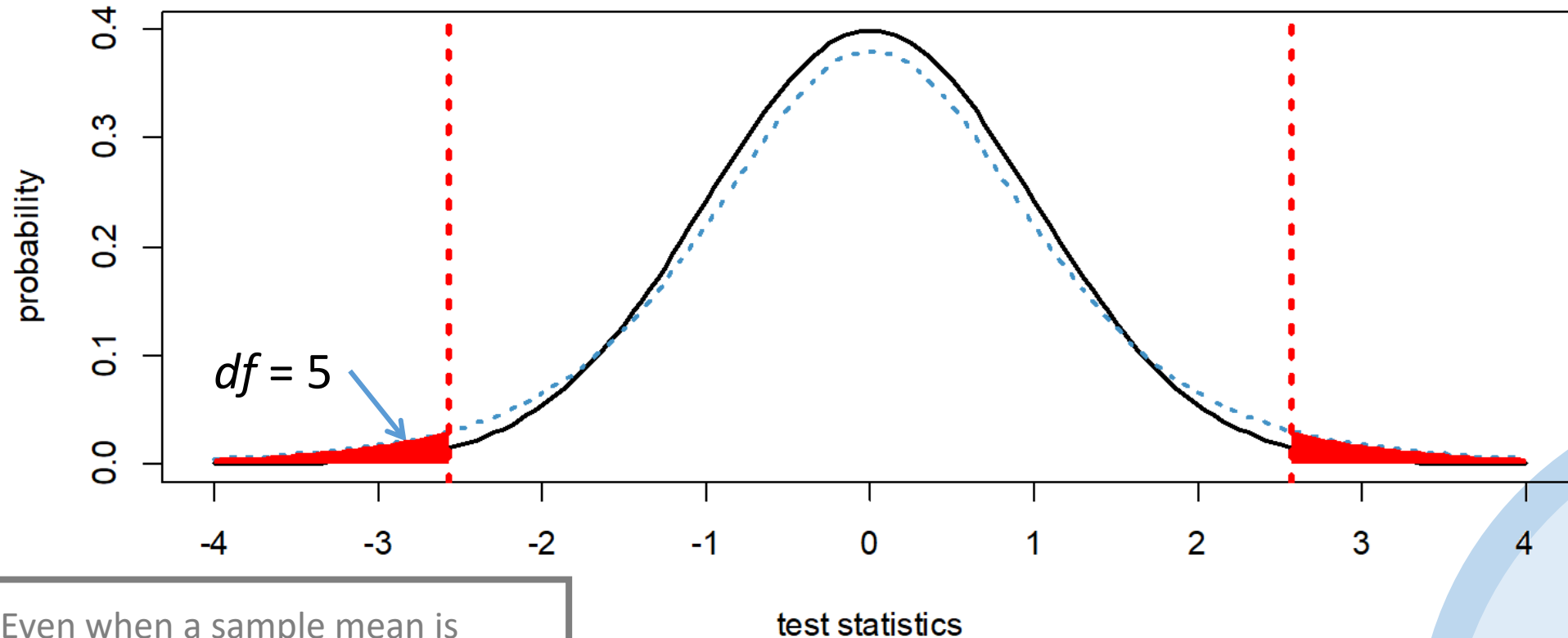


t distribution ($df = 5$)



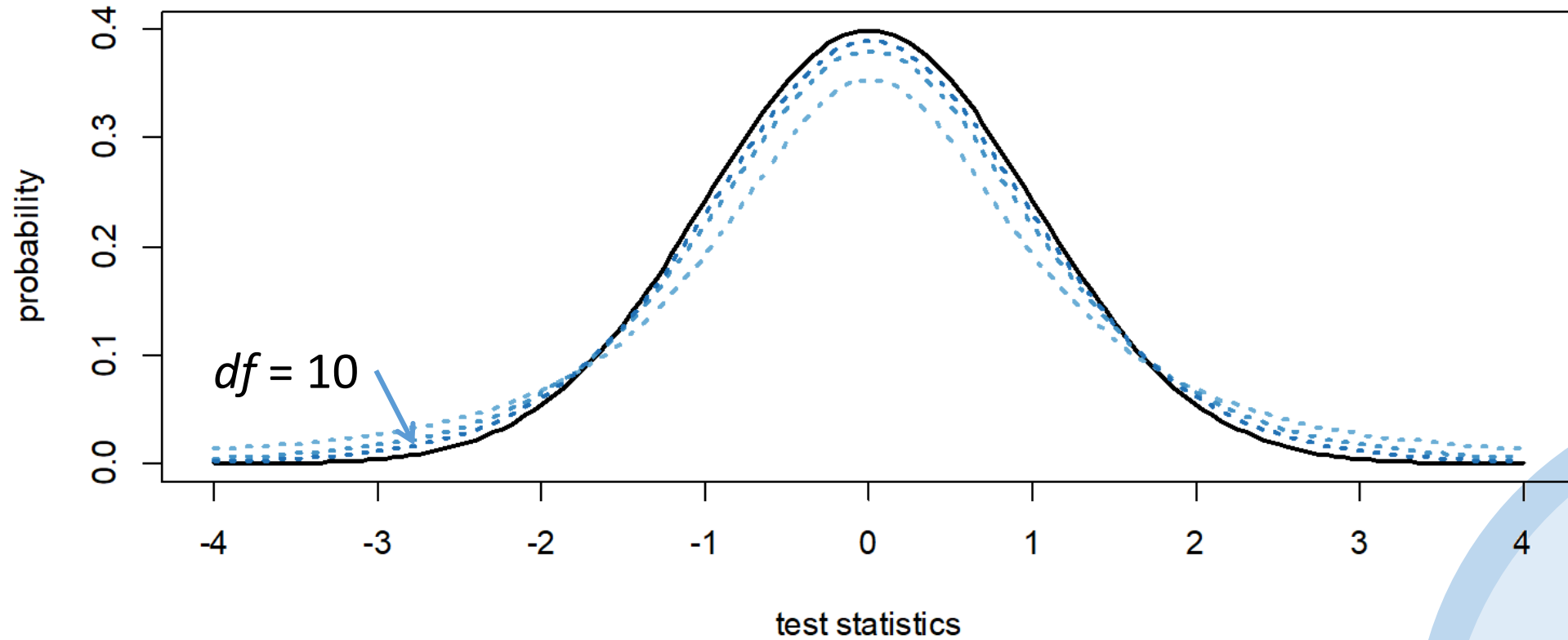
Estimation of the mean is less precise,
compared to the case when you have
the full distribution

t distribution ($df = 5$)



Even when a sample mean is 2.5 SE away from the null mean, you cannot be too confident that they are statistically different ($\alpha = 0.05$)

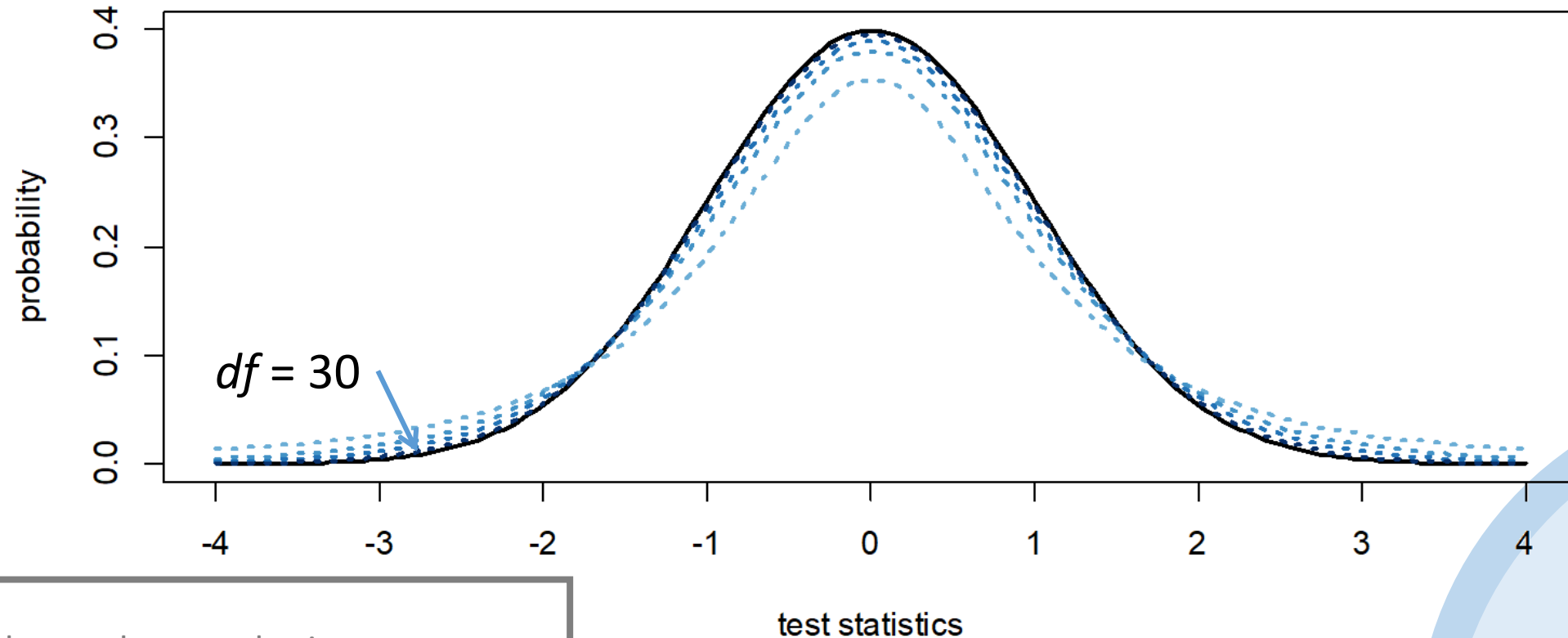
t distribution ($df = 10$)



t distribution ($df = 30$)

At $n = 30$:

- the sampling distribution is approaching normal (regardless of the population distribution)
- the t distribution ($df = 30$) is also approaching normal



The larger the sample size you use to estimate the population distribution, the closer it approaches normality

One-sample Z-test

When to use the test?	When population variance is known
Test Statistic & Formula	$Z_{Sample} = \frac{M_{treatment} - \mu_{baseline}}{\sigma_M}$
degrees of freedom (df)	---
Population Standard Deviation [^]	$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$
Standard Error [#]	$\sigma_M = \frac{\sigma}{\sqrt{n}}$
Quantile ↑↓ Percentile	Quantile → Percentile: <code>pnorm(q, mean = 0, sd = 1)</code> Percentile → Quantile: <code>qnorm(p, mean = 0, sd = 1)</code>
Confidence Interval	$M_{treatment} \pm Z_{\frac{\alpha}{2}} \times \sigma_M$

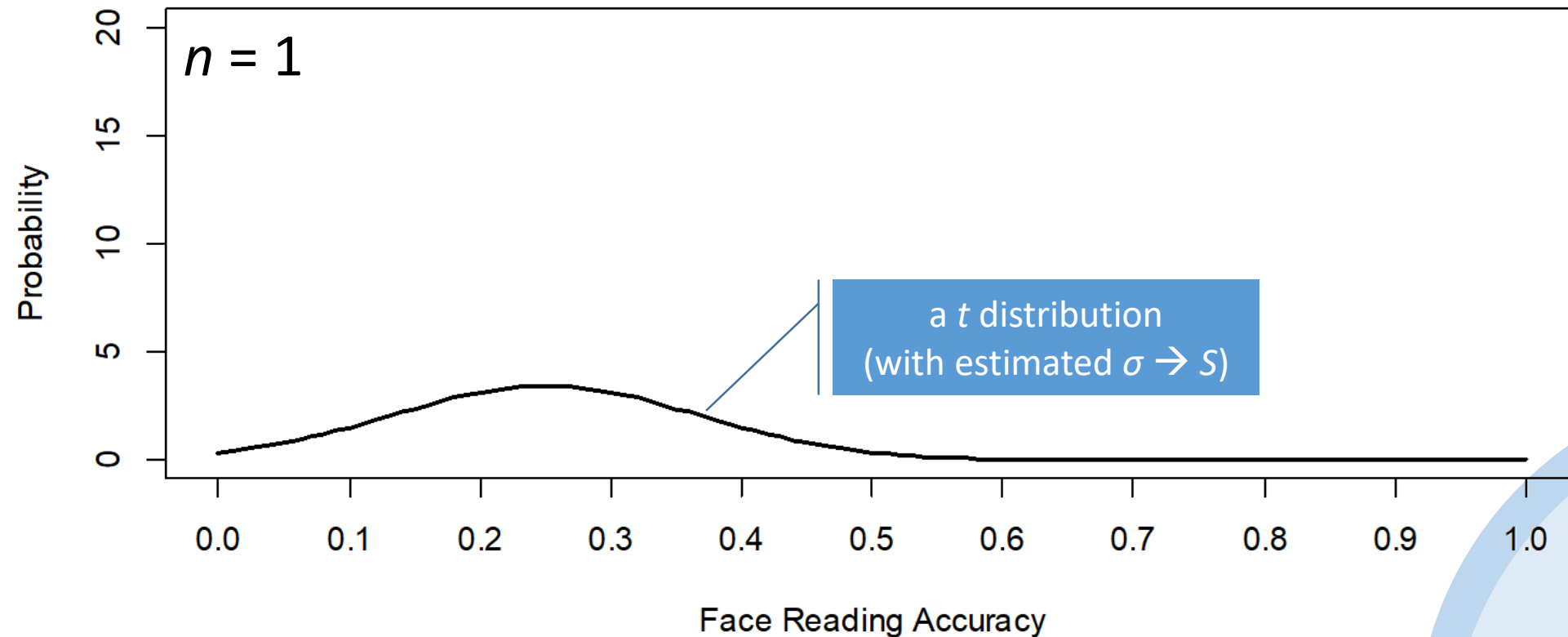
One-sample Student's t-test

When to use the test?	When population variance is unknown
Test Statistic & Formula	$t_{Sample} = \frac{M_{treatment} - \mu_{baseline}}{S_M}$
degrees of freedom (df)	$n - 1$
Population Standard Deviation [^]	Population standard deviation is estimated by: $S = \sqrt{\frac{\sum (X - M)^2}{n - 1}} = \sqrt{\frac{SS}{df}} \quad \text{sd}(X)$
Standard Error [#]	$S_M = \frac{S}{\sqrt{n}}$
Quantile ↑↓ Percentile	Quantile → Percentile: <code>pt(q, df)</code> Percentile → Quantile: <code>qt(p, df)</code>
Confidence Interval	$M_{treatment} \pm t_{\frac{\alpha}{2}, df} \times S_M$

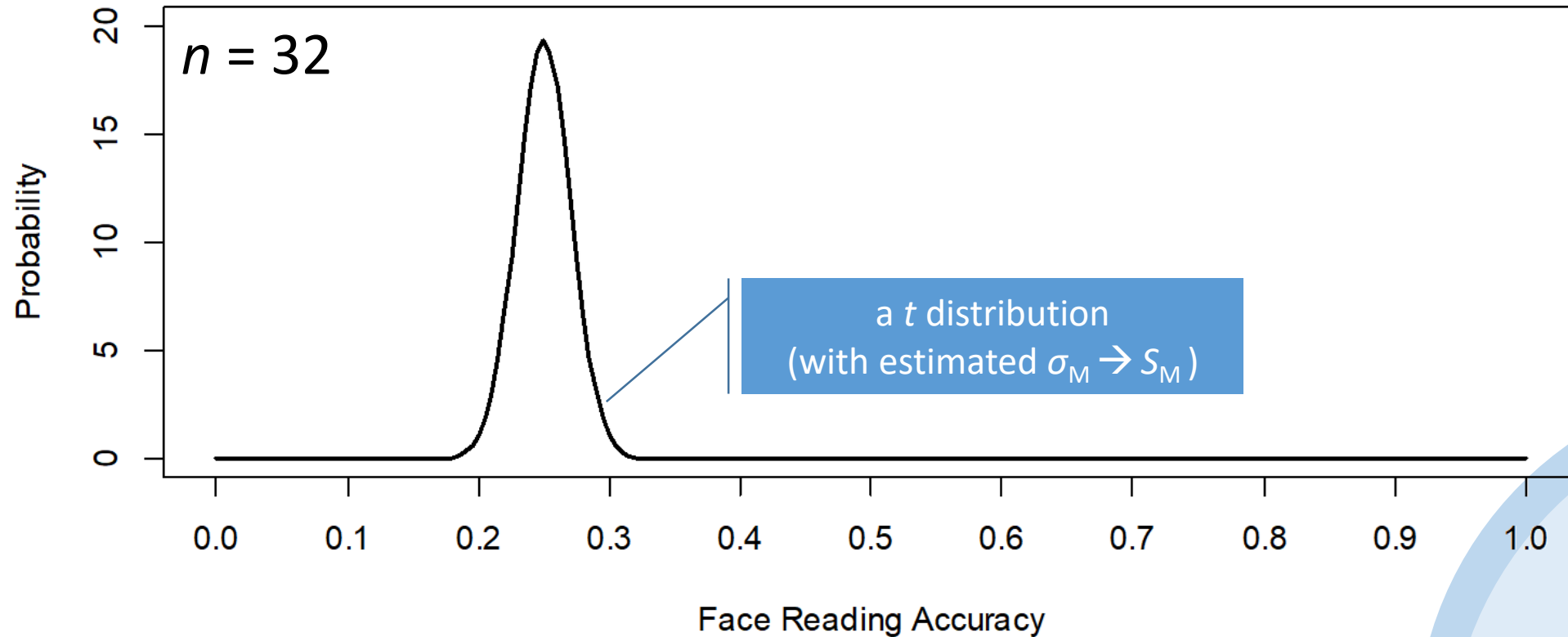
Note[^]: μ_M is the mean to be tested against the sample and M is the sample mean.

Note[#]: σ_M Standard Error is the Standard Deviation of *the distribution of sample means*.

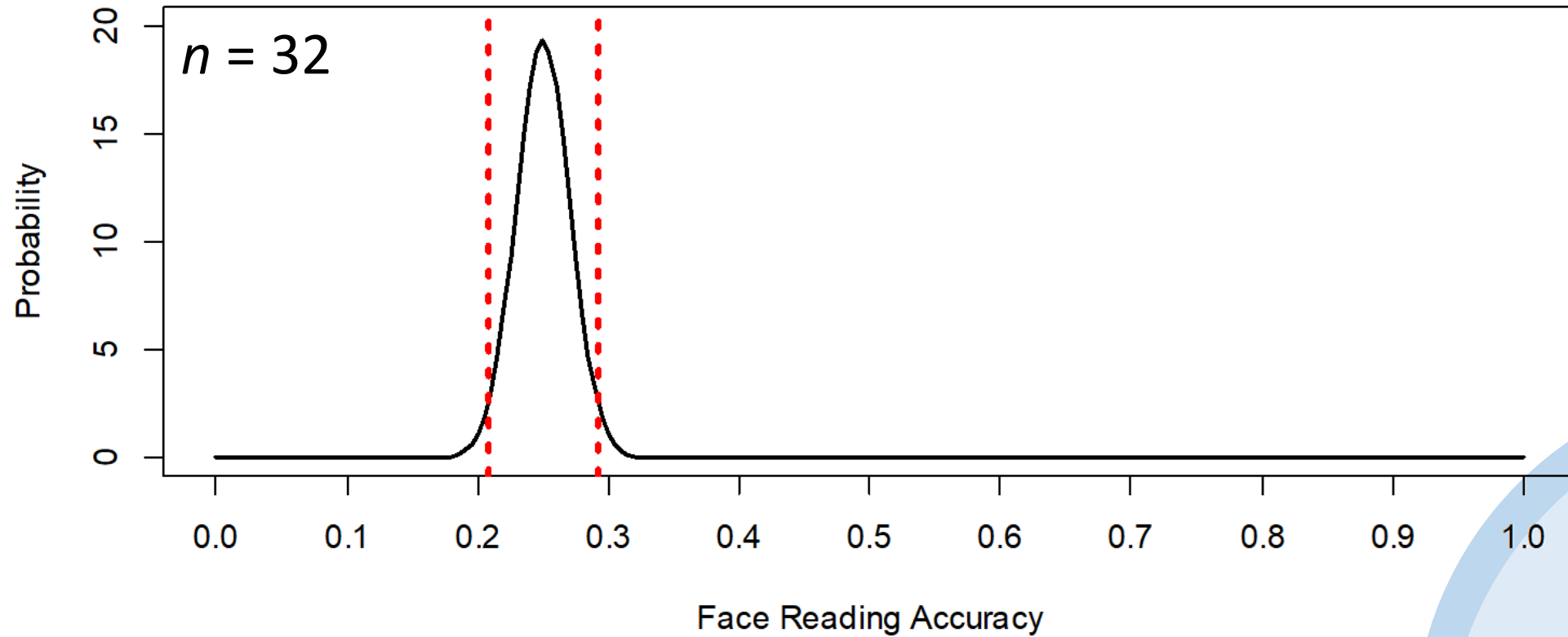
Population distribution: face reading experiment



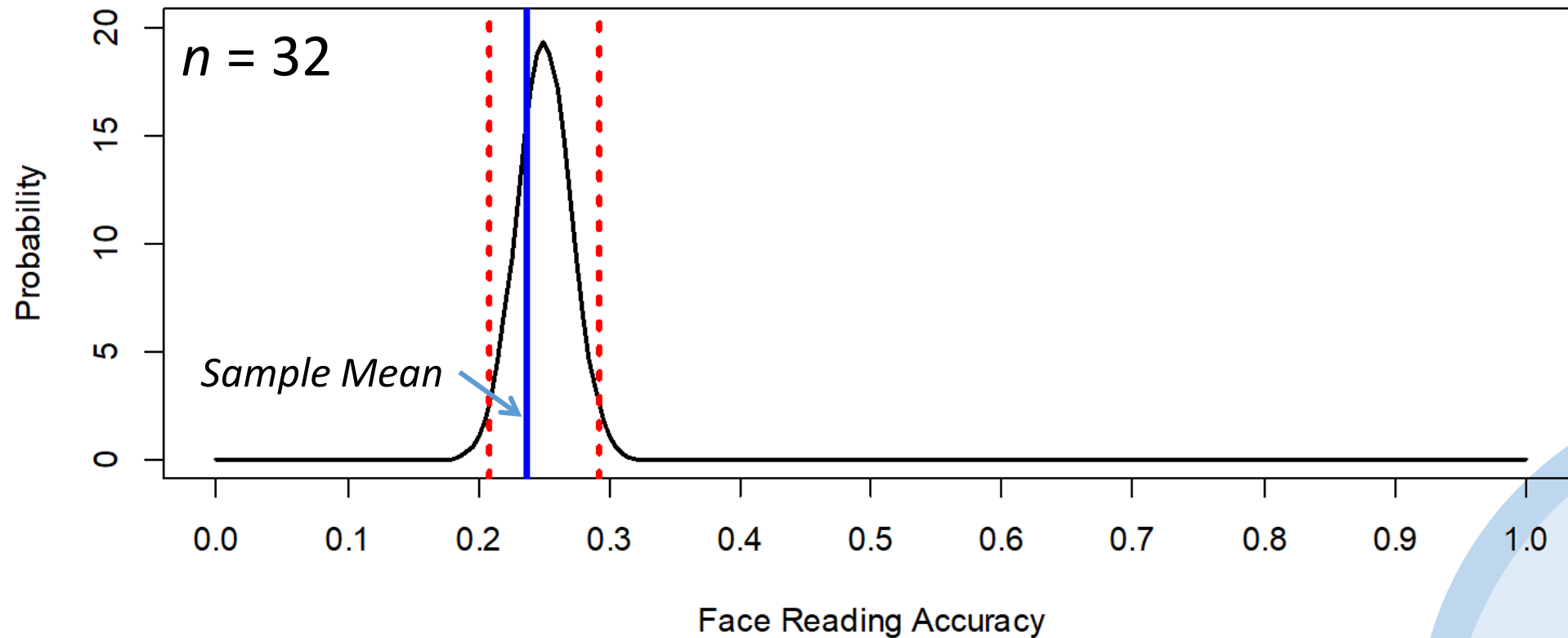
Sampling distribution



Sampling distribution



Sampling distribution



Hypothesis Testing – Step-by-step

- Step 0: Gather all the information you need, check the assumptions

- $\mu_{\text{baseline}} =$
 $\sigma = \text{unknown} \rightarrow \text{estimated with } S =$
 $M_{\text{treatment}} =$; $n =$; $df = n - 1$
kurtosis of population

- Step 1: Set up the hypotheses

- H_0 (null hypothesis)
- H_1 (Research hypothesis)

- $H_0:$
 $H_1:$

- Step 2: Calculate the test statistic

- $t_{\text{sample}} = \frac{M_{\text{treatment}} - \mu_{\text{baseline}}}{S_M}$

- Step 3: Find the probability associated with the test statistic
(i.e., the probability of getting such a score or a more extreme score x 2, given H_0 is true)

- $p(|t_{\text{sample}}| | H_0) =$

For a two-tailed test and t_{sample} on the left

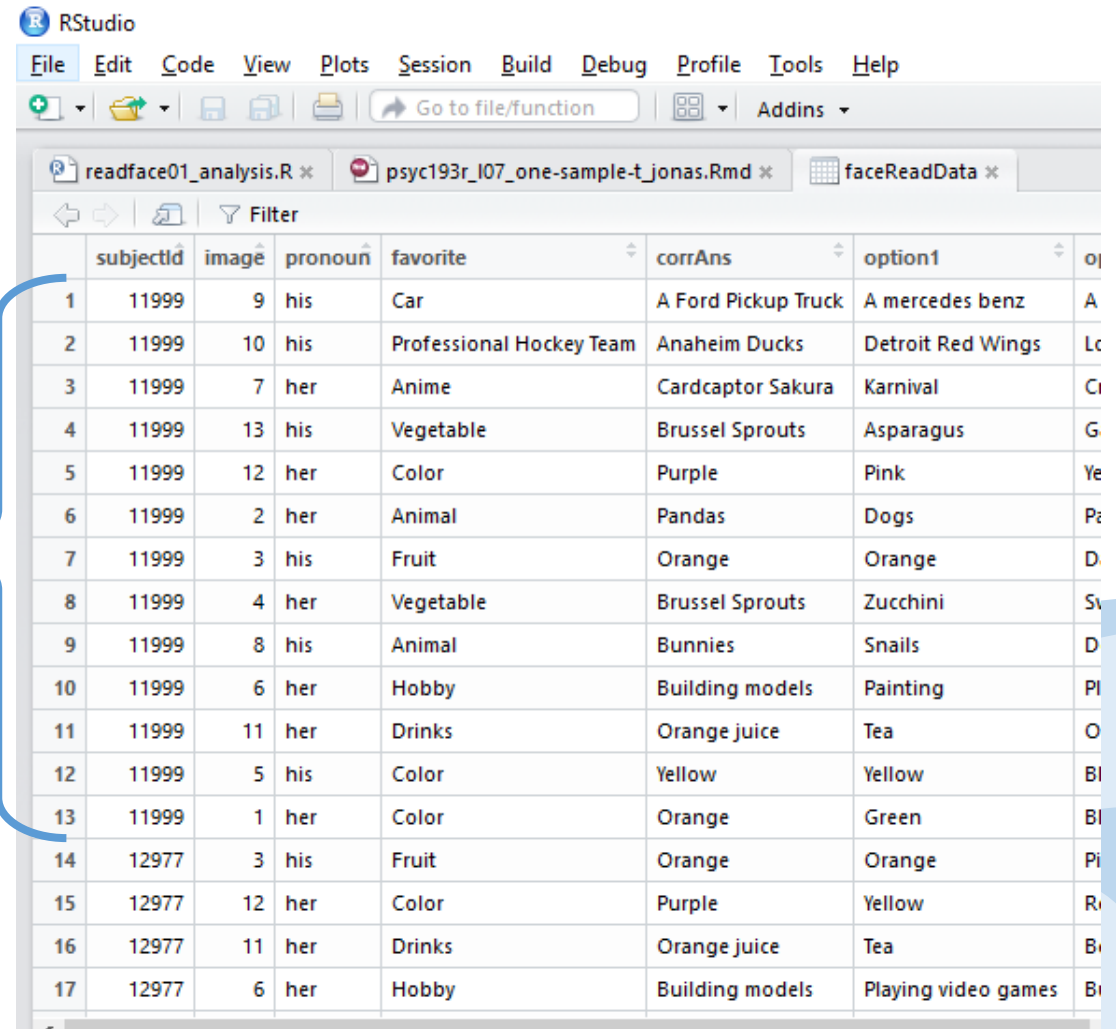
- Step 4: Decision & Conclusion
(compare p against α)

Import Data

- Download the data file:
https://psyc193r.ucsd.edu/data/psyc193r_readface_data.csv
- Tell R where your data file is:
Session → Set Working Directory → Choose Directory
`setwd()`
- Check the *comma separated values* (csv) file with MS Excel to ensure correct format
- Import the file into R as a data frame, saving it to “faceReadData” variable
`... = read.csv("file_name.csv")`

Data Structure

- The data have a "long format":
 - Each column represents a variable
 - Each row represents an entry (each trial)
- The variable we care the most are "subjectId", & "correct"
 - subjectId: a number that represents a participant (each should have 13 rows)
 - correct: denotes if a participant is correct on a particular trial



The screenshot shows the RStudio interface with a data table named 'faceReadData' open in the Environment pane. The table has 8 columns: 'subjectId', 'image', 'pronoun', 'favorite', 'corrAns', 'option1', and 'option2'. The first 13 rows correspond to subjectId 11999, and the next 4 rows correspond to subjectId 12977. A blue bracket on the left side of the table highlights the first 13 rows, indicating that each participant has 13 trials.

	subjectId	image	pronoun	favorite	corrAns	option1	option2
1	11999	9	his	Car	A Ford Pickup Truck	A mercedes benz	A
2	11999	10	his	Professional Hockey Team	Anaheim Ducks	Detroit Red Wings	Lc
3	11999	7	her	Anime	Cardcaptor Sakura	Karnival	Cr
4	11999	13	his	Vegetable	Brussel Sprouts	Asparagus	G
5	11999	12	her	Color	Purple	Pink	Ye
6	11999	2	her	Animal	Pandas	Dogs	Pa
7	11999	3	his	Fruit	Orange	Orange	D
8	11999	4	her	Vegetable	Brussel Sprouts	Zucchini	Sw
9	11999	8	his	Animal	Bunnies	Snails	D
10	11999	6	her	Hobby	Building models	Painting	Pl
11	11999	11	her	Drinks	Orange juice	Tea	O
12	11999	5	his	Color	Yellow	Yellow	Bl
13	11999	1	her	Color	Orange	Green	Bl
14	12977	3	his	Fruit	Orange	Orange	Pi
15	12977	12	her	Color	Purple	Yellow	R
16	12977	11	her	Drinks	Orange juice	Tea	Br
17	12977	6	her	Hobby	Building models	Playing video games	Br

Data Cleaning

- Convert `subjectId` into a factor
it means `subjectId` are nominal variables (they're just names)
Otherwise R would treat the variable as numerics
`faceReadData$subjectId = factor(faceReadData$subjectId)`
- Coerce / Make sure `correct` is a variable
so that we can do some computations with it
`faceReadData$correct = as.numeric(faceReadData$correct)`



Data Cleaning

- What we want is not exactly trial-level data, instead, we want to know the mean accuracy of each individual

We want to make a new data frame with two variables (columns), and save it to a new variable ("faceRead")

- subjectId
- accuracy

the data set

```
... = aggregate(correct ~ subjectId, data=faceReadData, FUN = mean)
```

group the scores "correct" by the same "subjectId"

use this function for the grouping

- Change the name of the aggregated variable to "accuracy"

```
names(faceRead)[2] = "accuracy"
```

name of the 2nd variable in "faceRead"