Midterm Review

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

November 3, 2018

Exam Details

- **When:** Monday October 29, 11:30 am. McMED 504 (McIntyre Medical Building)
- This is a 2 hour, open book exam.
- Calculators are permitted. Cellular phones are not permitted.
- The exam is out of 100. Write down all your answers in the provided booklet.
- Provide units and state your assumptions when applicable.
- If a question requires use of the z or t probabilites/quantiles, write the corresponding R code instead. Some commonly used quantiles are provided.

Topics to be covered

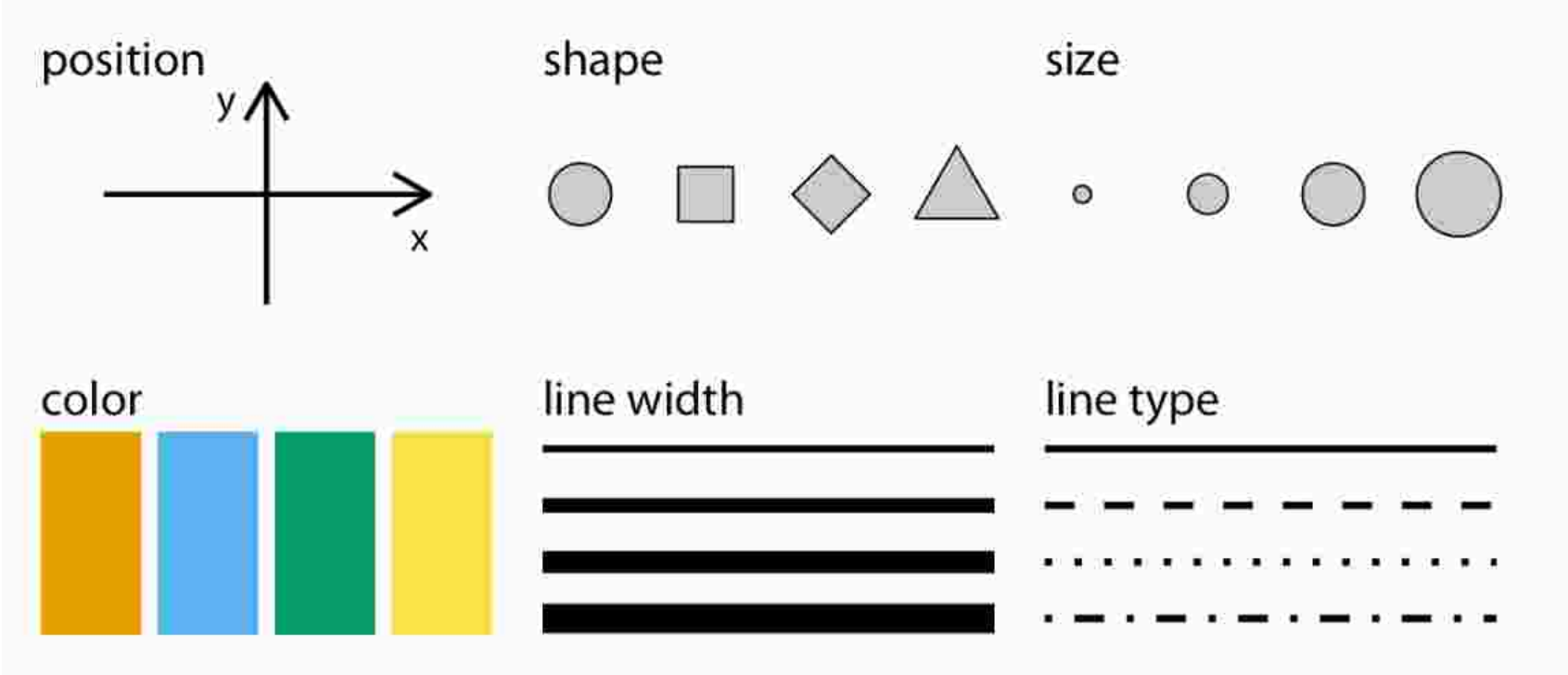
Topics to be covered

1. Data visualization (1)
2. Descriptive statistics (4)
3. Sampling Distributions, CLT, Confidence intervals and p-values
4. One sample mean (5)
5. One sample proportion (6)
6. Power (1)
7. Bootstrap (1)

Data visualization

Aesthetics

- Aesthetics



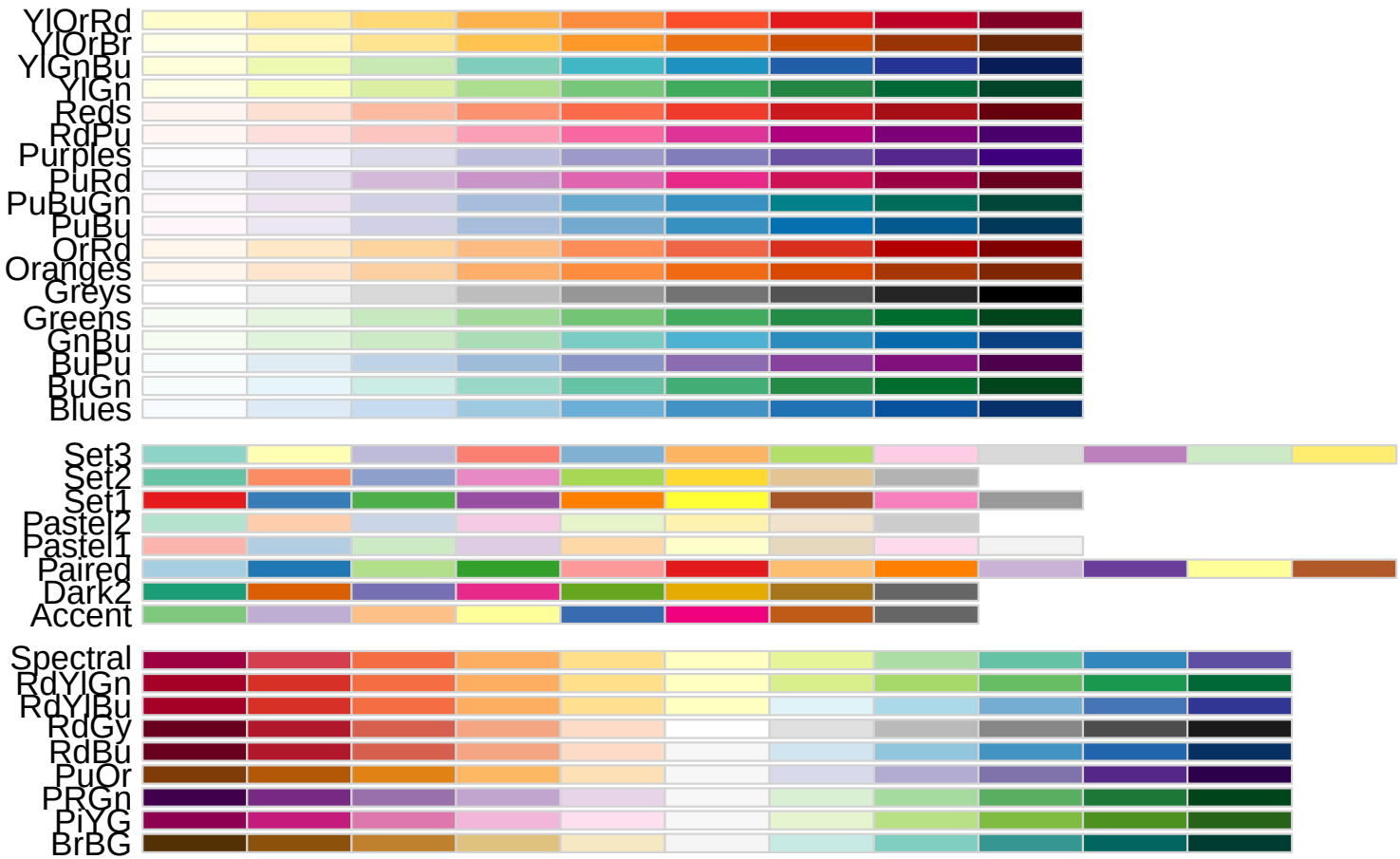
- Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can only represent discrete data (shape, line type)

Variable Types

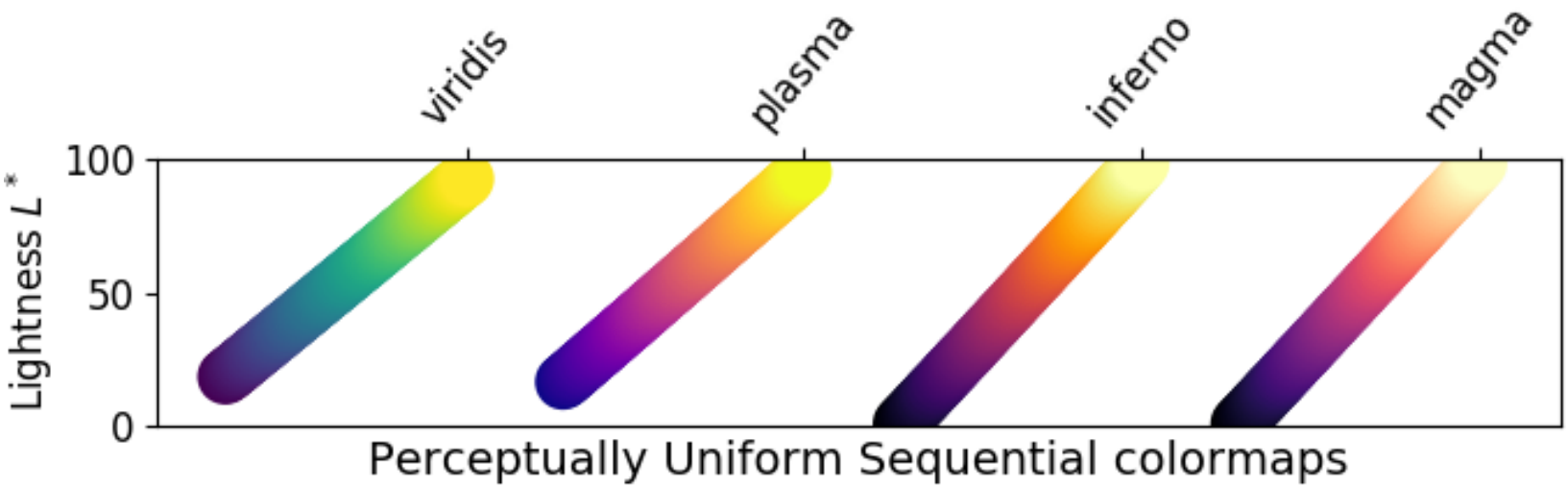
- quantitative/numerical continuous (1.3, 5.7, 83, 1.5×10^{-2})
- quantitative/numerical discrete (1,2,3,4)
- qualitative/categorical unordered (dog, cat, fish)
- qualitative/categorical ordered (good, fair, poor)

Color Palettes: Cynthia Brewer

```
pacman::p_load(RColorBrewer)
RColorBrewer::display.brewer.all()
```



Color Palettes: viridis



Descriptive statistics

Descriptive statistics

- Boxplots, histograms, density plot
- IQR, median, mode, mean, min, max
- Q1, Q3
- Skewness (long left/right tail)

Standard error (SE) of a sample statistic

- Recall: When we are talking about the variability of a **statistic**, we use the term **standard error** (not standard deviation). The standard error of the sample mean is $\sigma/\sqrt{n}$.

Remark 1 (SE vs. SD)

In quantifying the instability of the sample mean ($\bar{y}$) statistic, we talk of SE of the mean (SEM)

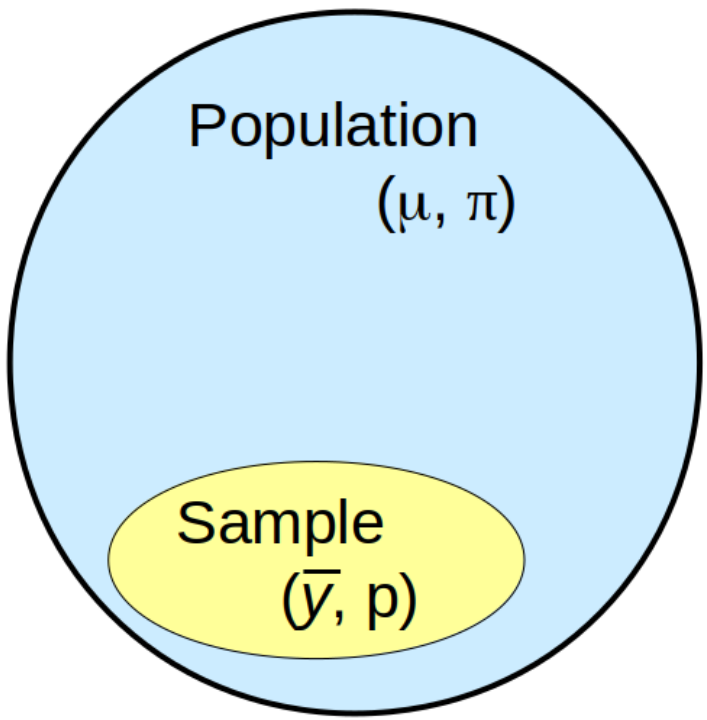
SE($\bar{y}$) describes how far $\bar{y}$ could (typically) deviate from μ;

SD(y) describes how far an individual y (typically) deviates from μ (or from $\bar{y}$).

Sampling Distributions, CLT, Confidence Intervals and p-values

Parameters, Samples, and Statistics

- **Paramter**: An unknown numerical constant pertaining to a population/universe, or in a statistical model.
  - μ: population mean π: population proportion
- **Statistic**: A numerical quantity calculated from a sample. The empirical counterpart of the parameter, used to *estimate* it.
  - $\bar{y}$: sample mean p: sample proportion



Samples must be random

- The validity of inference will depend on the way that the sample was collected. If a sample was collected badly, no amount of statistical sophistication can rescue the study.

- Samples should be **random**. That is, there should be no systematic set of characteristics that is related to the scientific question of interest that causes some people to be more likely to be sampled than others. The simplest type of randomization selects members from the population with equal probability (a uniform distribution).

- Do not cheat by
  - Taking 5 people from the *same* household to estimate
    - proportion of Québécois who don't have a family doctor
    - who saw a medical doctor last year
    - average rent
  - Sampling the depth of the ocean *only around Montreal* to estimate
    - proportion of Earth's surface covered by water

Sampling Distributions

Definition 1 (Sampling Distribution)

- *The sampling distribution of a statistic is the distribution of values taken by the statistic in* ***all possible samples of the same size*** *from the same population.*

- *The standard deviation of a sampling distribution is called a* ***standard error***

Sampling Distributions

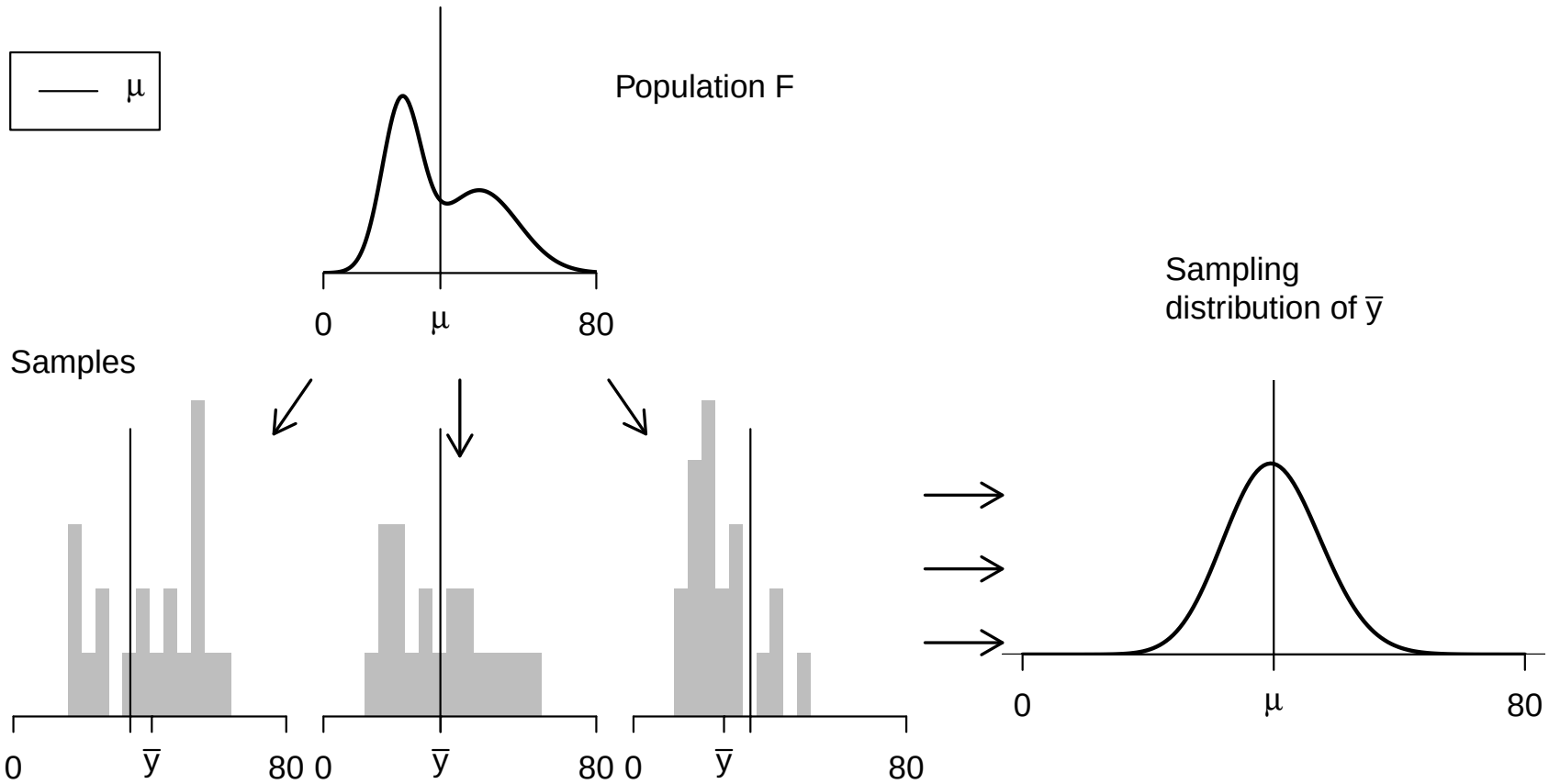


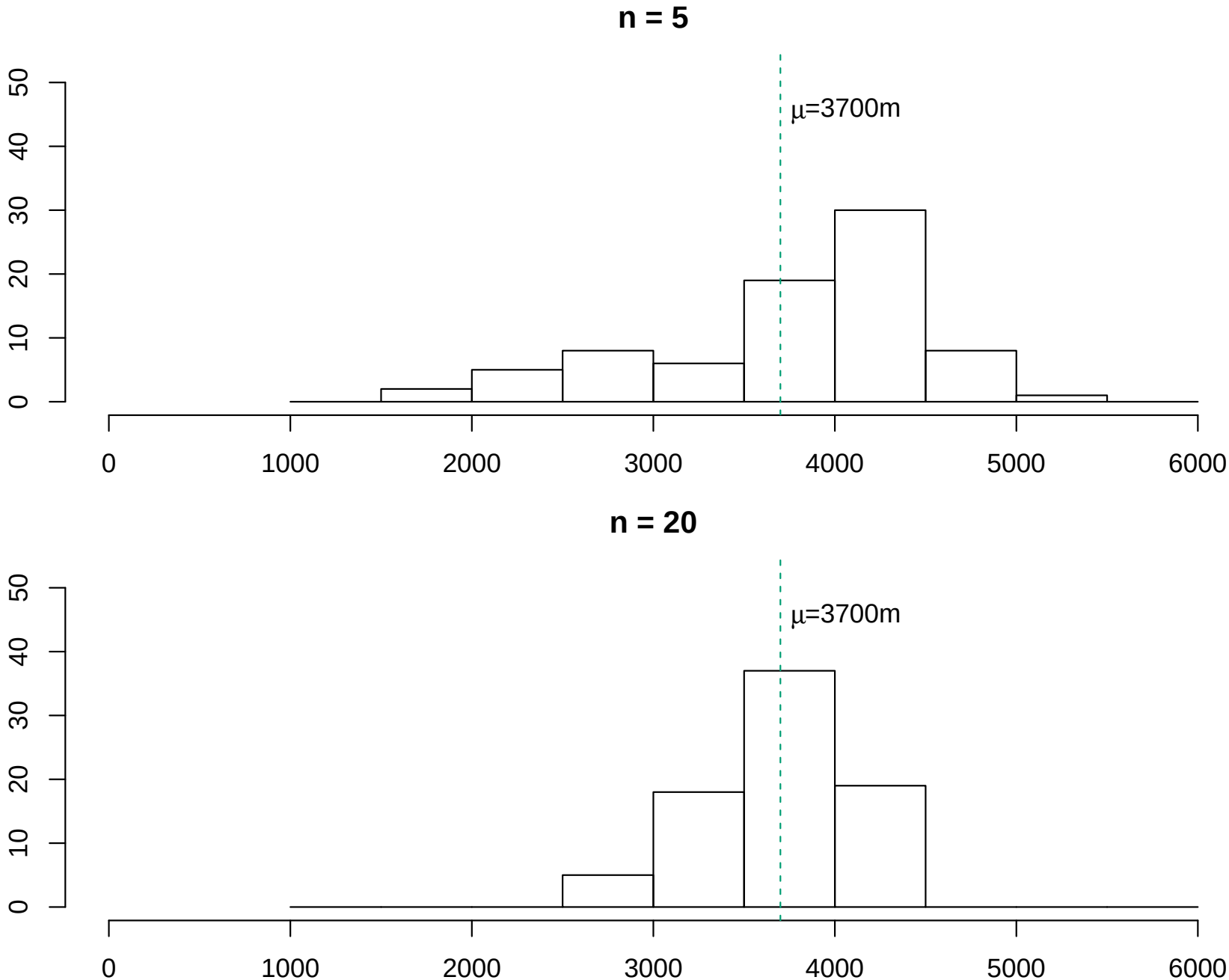
Fig.: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Why are sampling distributions important?

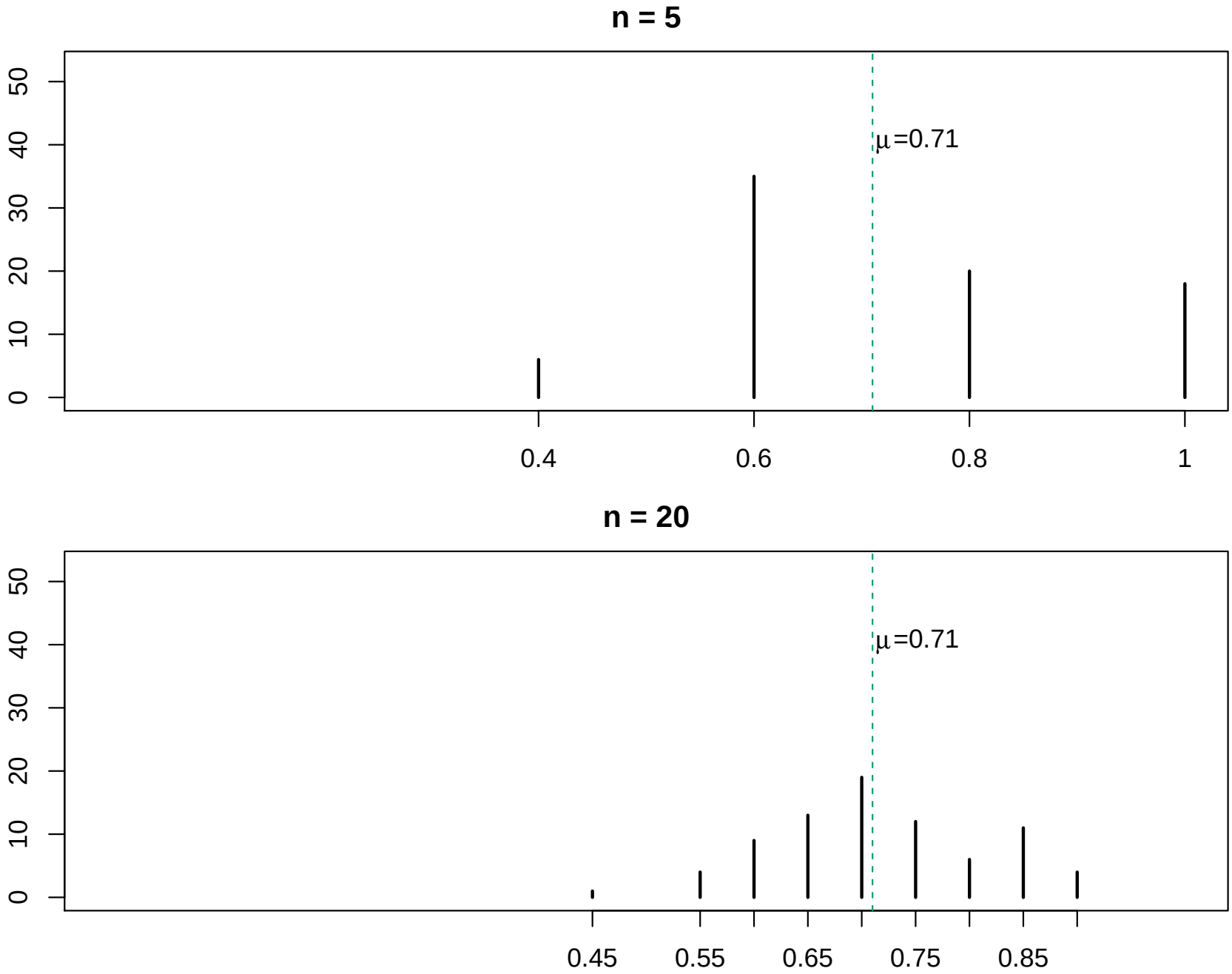
- They tell us how far from the target (true value of the parameter) our statistical *shot* at it (i.e. the statistic calculated form a sample) is likely to be, or, to have been.

- Thus, they are used in confidence intervals for parameters. Specific sampling distributions (based on a null value for the parameter) are also used in statistical tests of hypotheses.

Sampling distribution: mean depth of the ocean



Sampling distribution: proportion covered by water

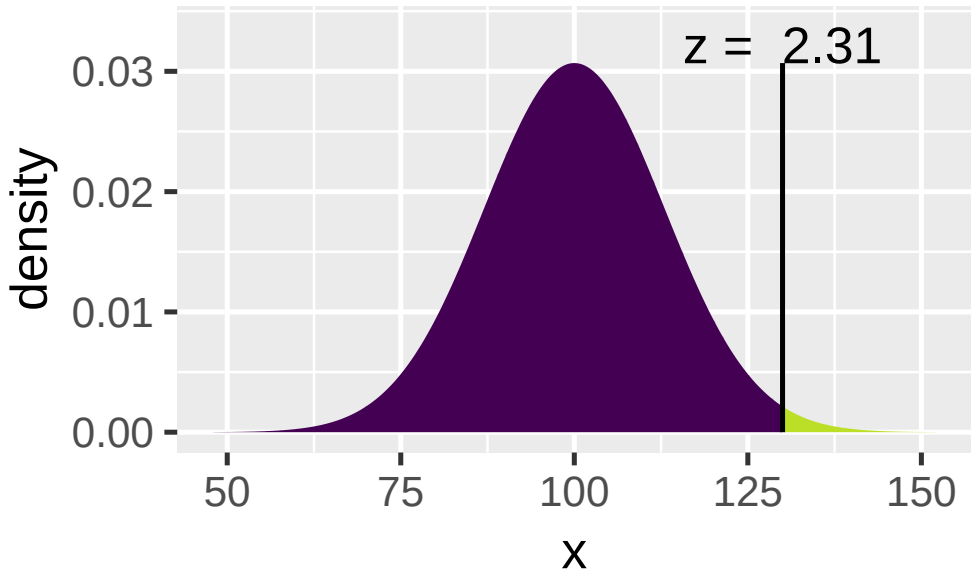


Normal Distribution: For probabilities we use *pnorm*

```
stats::pnorm(q = 130, mean = 100, sd = 13)
```

```
## [1] 0.9894919
```

```
mosaic::xpnorm(q = 130, mean = 100, sd = 13)
```




```
## [1] 0.9894919
```

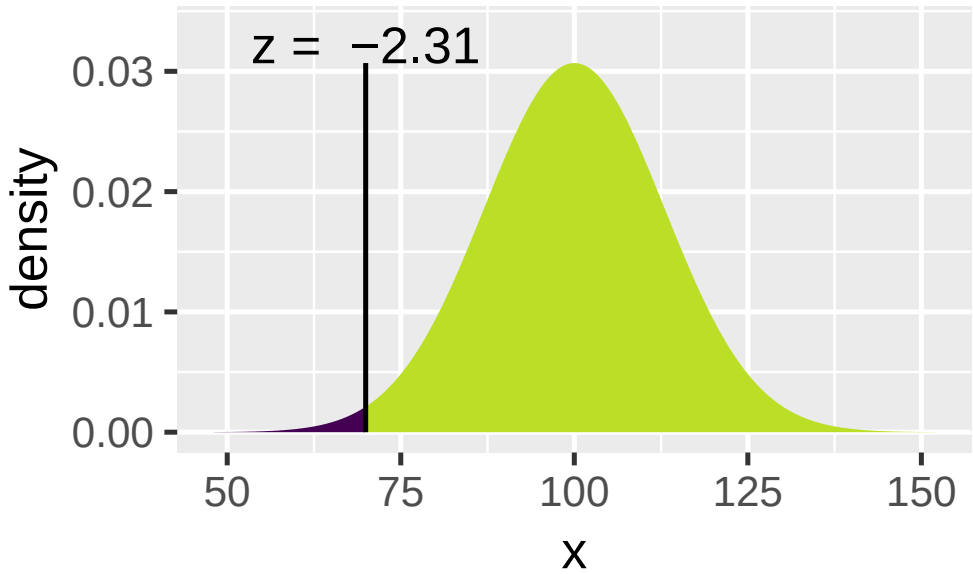
- `pnorm` returns the integral from $-\infty$ to q for a $\mathcal{N}(\mu, \sigma)$
- `pnorm` goes from *quantiles* (think Z scores) to probabilities

Normal Distribution: For quantiles we use *qnorm*

```
stats::qnorm(p = 0.0104, mean = 100, sd = 13)
```

```
## [1] 69.94926
```

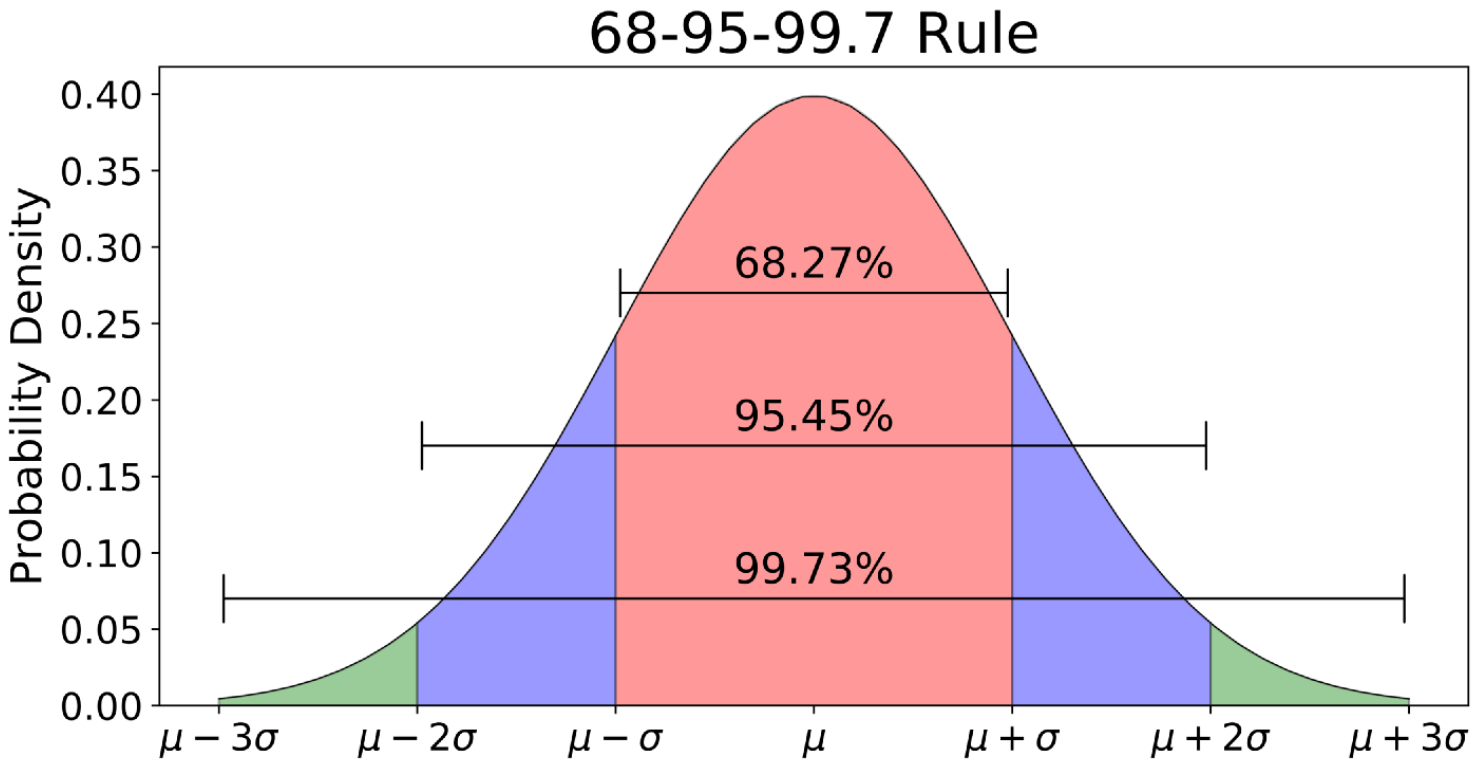
```
mosaic::xqnorm(p = 0.0104, mean = 100, sd = 13)
```




```
## [1] 69.94926
```

- `qnorm` answers the question: What is the Z-score of the *p*th percentile of the normal distribution?
- `qnorm` goes from *probabilities* to quantiles

Empirical Rule or 68-95-99.7% Rule



Quadruple the work, half the benefit

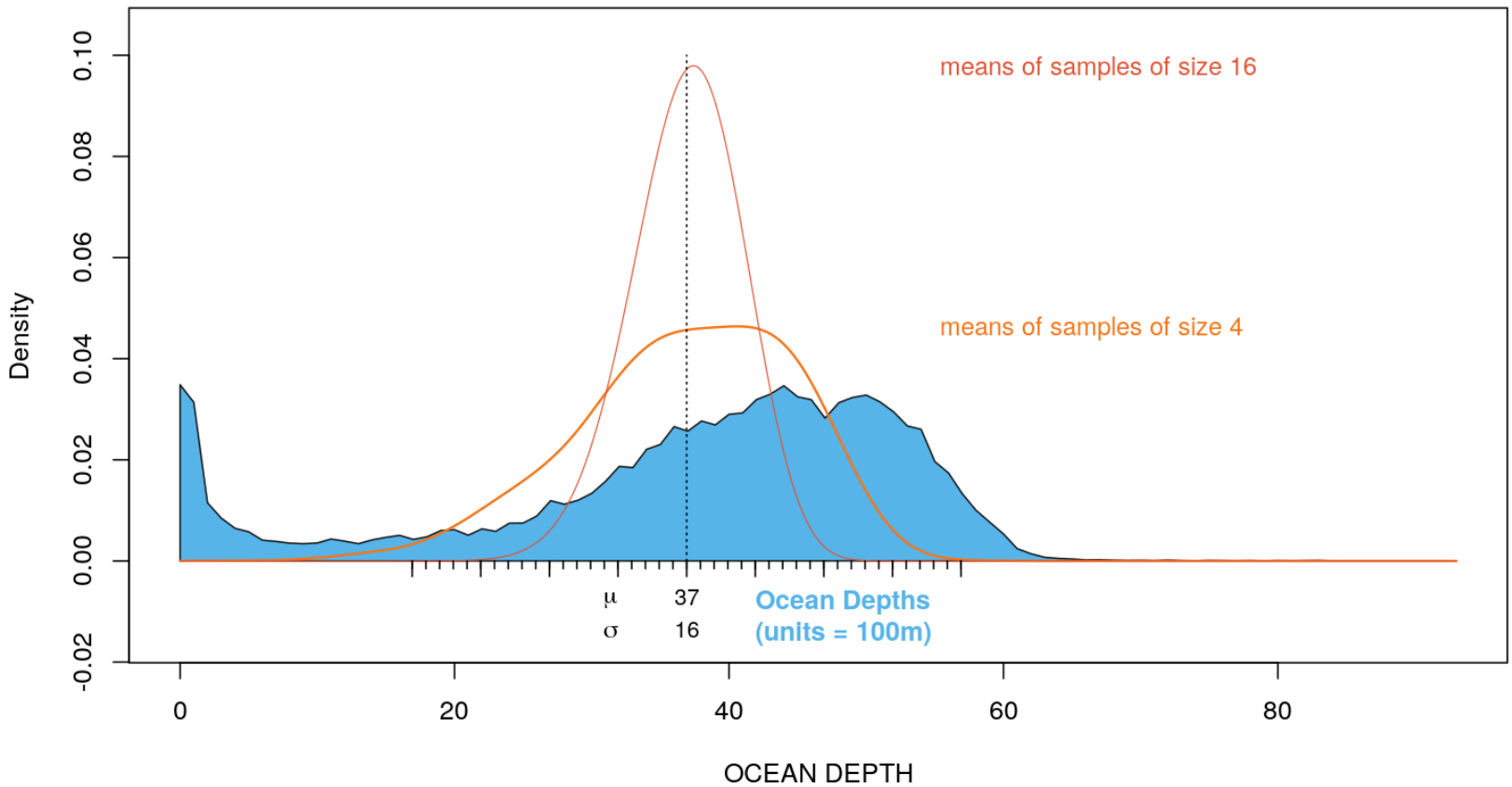


Fig.: When the sample size increases from 4 to 16, the spread of the sampling distribution for the mean is reduced by a half, i.e., the range is cut in half. This is known as the curse of the $\sqrt{n}$

The Central Limit Theorem (CLT)

- The sampling distribution of $\bar{y}$ is, for a large enough n, close to Gaussian in shape no matter what the shape of the distribution of individual Y values.
- This phenomenon is referred to as the CENTRAL LIMIT THEOREM
- The CLT applied also to a sample proportion, slope, correlation, or any other statistic created by aggregation of individual observations

Theorem 1 (Central Limit Theorem)

$$\text{if } Y \sim ???(\mu_Y, \sigma_Y), \text{ then}$$

$$\bar{y} \sim \mathcal{N}(\mu_Y, \sigma_Y/\sqrt{n})$$

Confidence Interval

Definition 2 (Confidence Interval)

A level C confidence interval for a parameter has two parts:

1. *An interval calculated from the data, usually of the form*

$$estimate \pm margin\ of\ error$$

   where the estimate is a sample statistic and the margin of error represents the accuracy of our guess for the parameter.

2. *A confidence level C, which gives the probability that the interval will capture the true parameter value in different possible samples. That is, the confidence level is the success rate for the method*

Confidence Interval: A simulation study

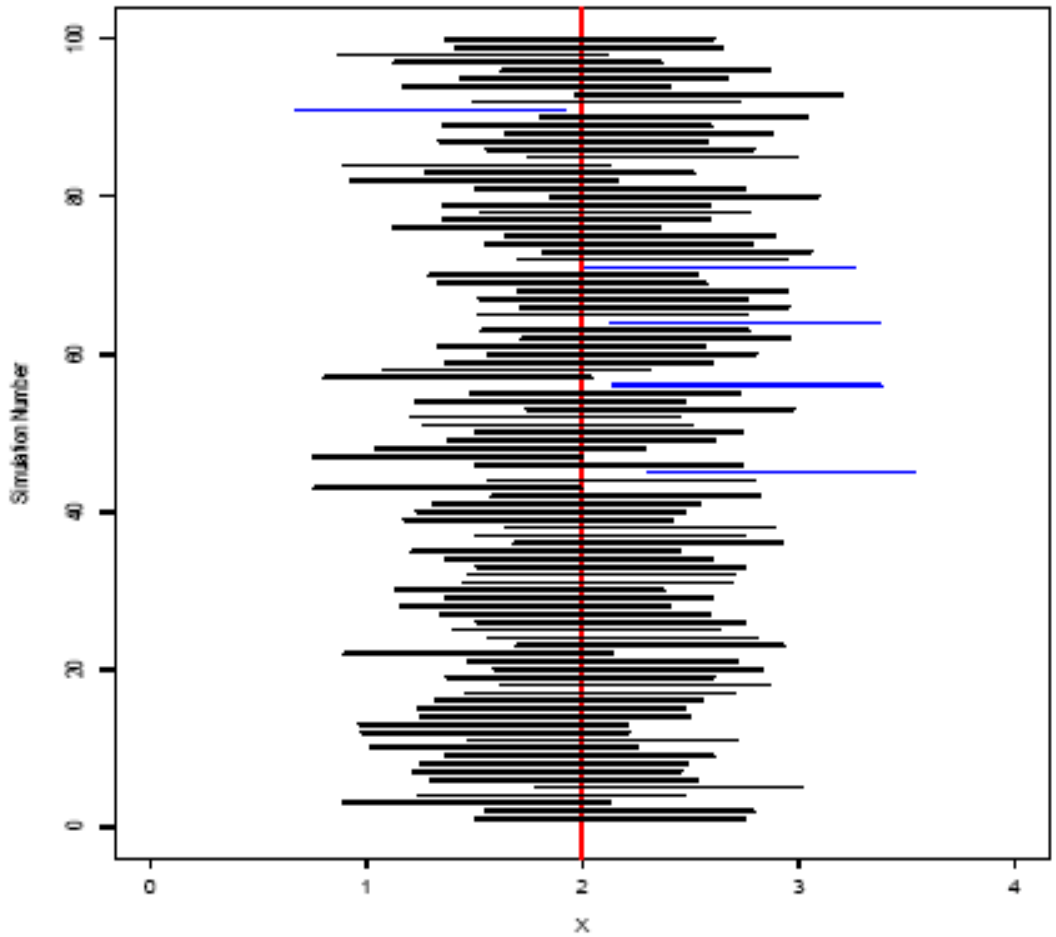


Fig.: True parameter value is 2 (red line). Each horizontal black line represents a 95% CI from a sample and contains the true parameter value. The blue CIs do not contain the true parameter value. 95% of all samples give an interval that contains the population parameter.

Interpreting a frequentist confidence interval

- The confidence level is the success rate of the method that produces the interval.
- We don't know whether the 95% confidence interval from a particular sample is one of the 95% that capture θ (the unknown population parameter), or one of the unlucky 5% that miss.
- To say that we are 95% confident that the unknown value of θ lies between U and L is shorthand for "We got these numbers using a method that gives correct results 95% of the time."

68% Confidence interval using `qnorm`

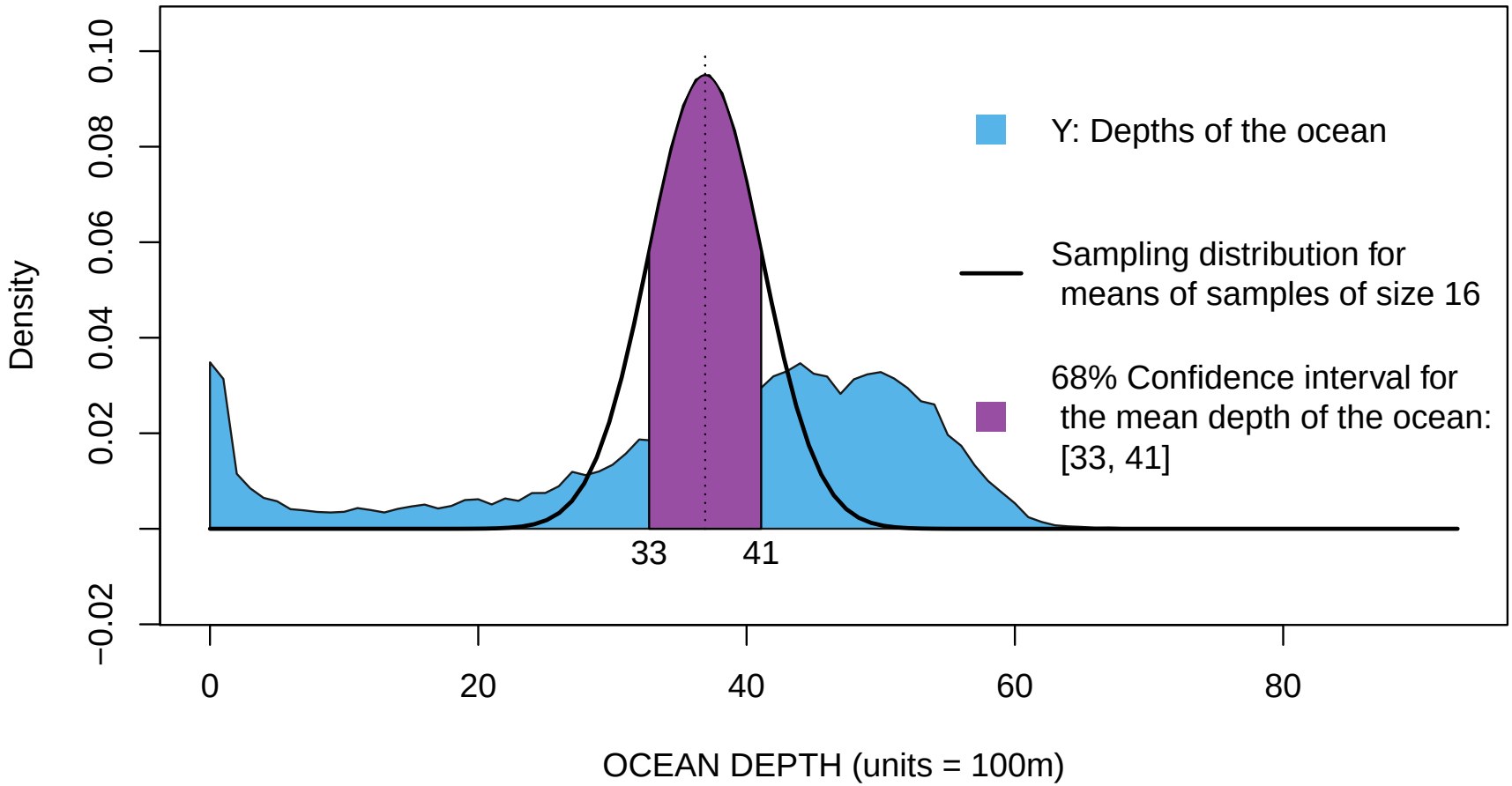


Fig.: 68% Confidence interval calculated using

```
qnorm(p = c(0.16,0.84), mean = 37, sd = 4.2)
```

95% Confidence interval using `qnorm`

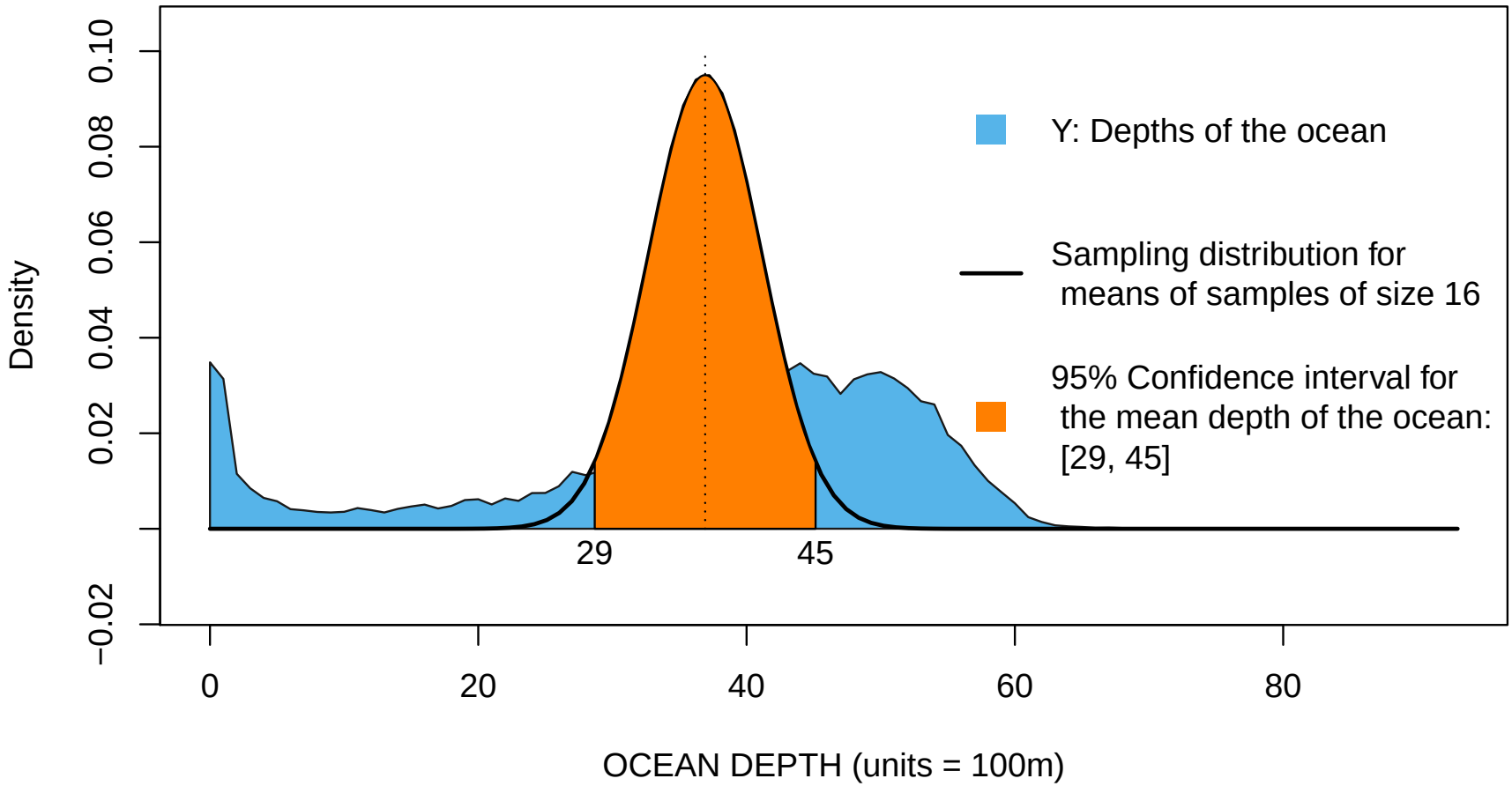


Fig.: 95% Confidence interval calculated using
`qnorm(p = c(0.025,0.975), mean = 37, sd = 4.2)`

Example: Inference for a single population mean

So what does the CI allow us to learn about μ??

- It tells us that if we repeated this procedure again and again (collecting a sample mean, and constructing a 95% CI), 95% of the time, the CI would *cover* μ.

- That is, with 95% probability, the *procedure* will include the true value of μ. Note that we are making a probability statement about the CI, not about the parameter.

- Unfortunately, we do not know whether the true value of μ is contained in the CI in the particular experiment that we have performed.

Bootstrap

Motivation for the Bootstrap

- The $\pm$ and `qnorm` methods to calculate a CI both require the CLT

Q: What happens if the CLT hasn't 'kicked in'?
Or you don't believe the CLT?

A: Bootstrap

The Bootstrap

Ideal world: known sampling distribution

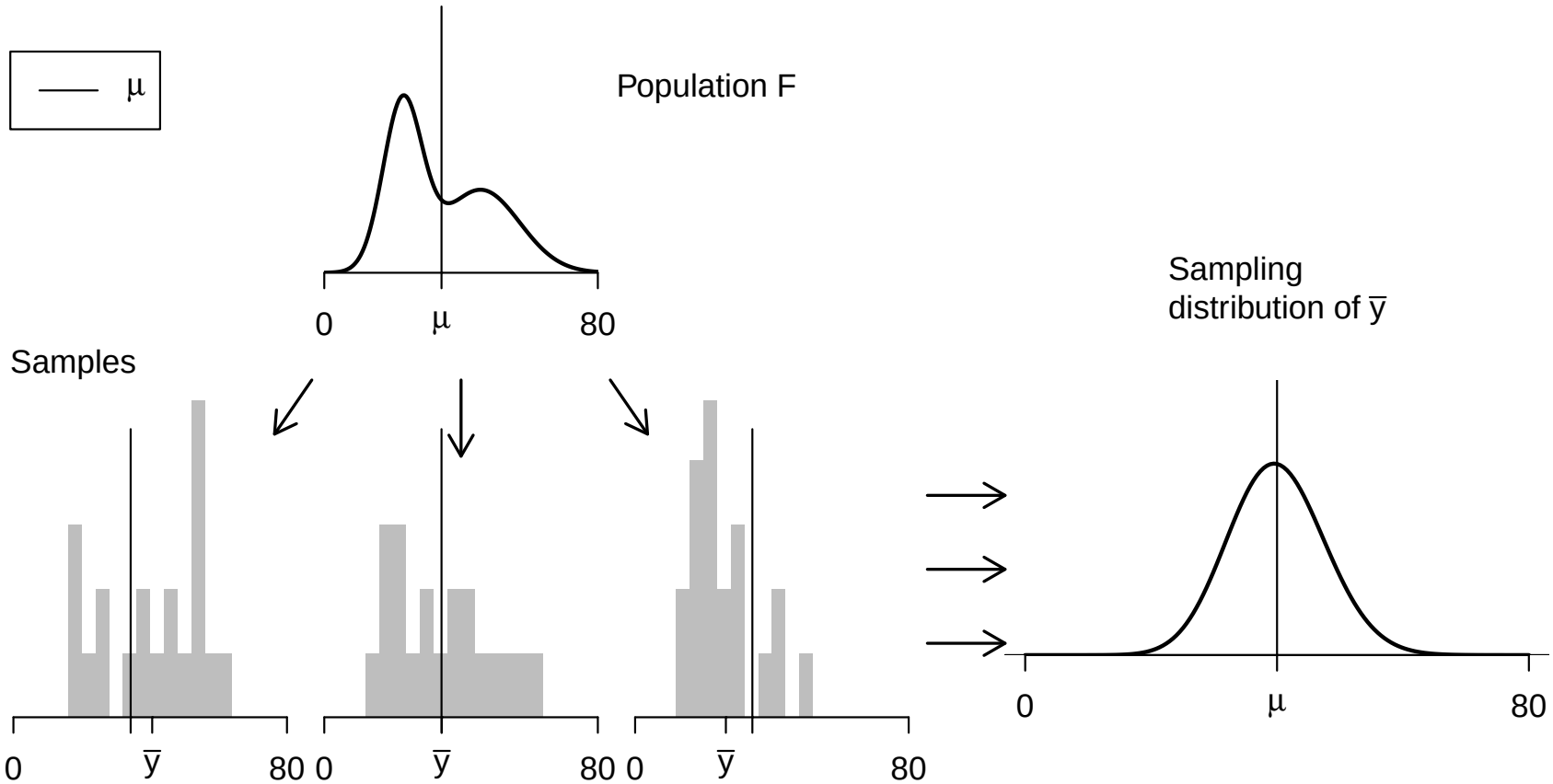


Fig.: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Reality: use the bootstrap distribution instead

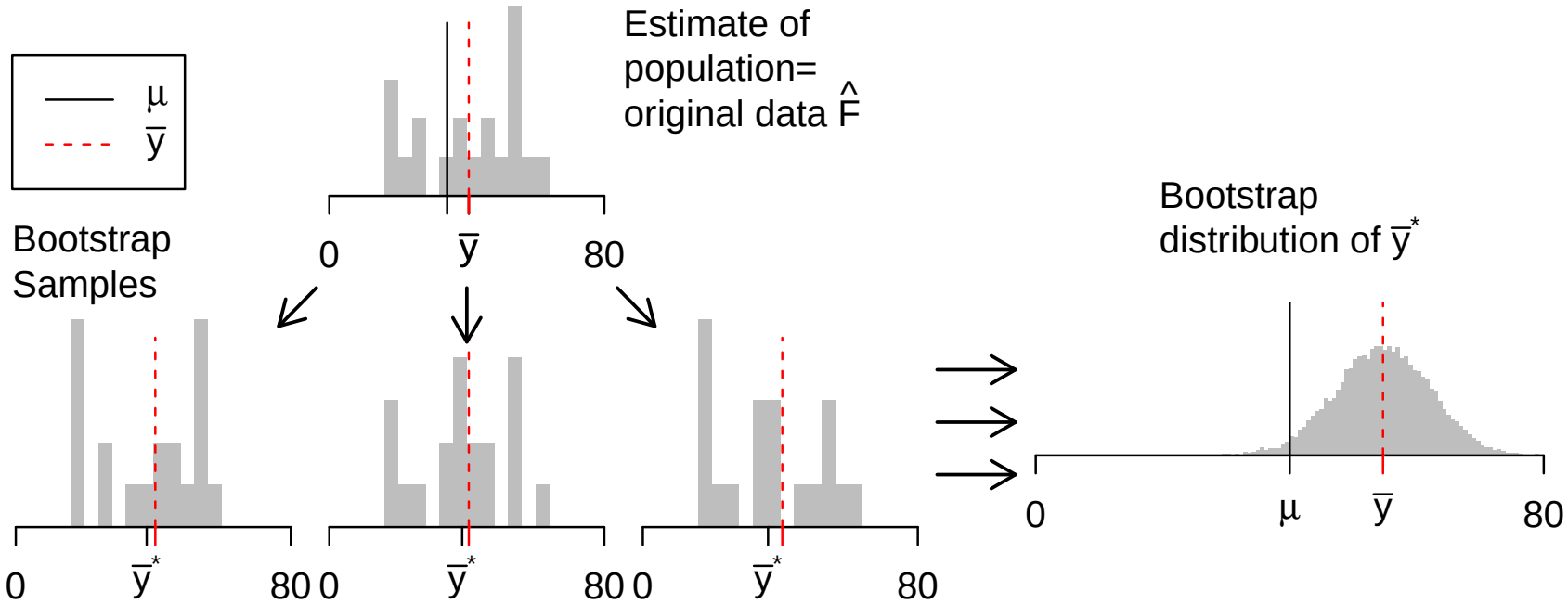
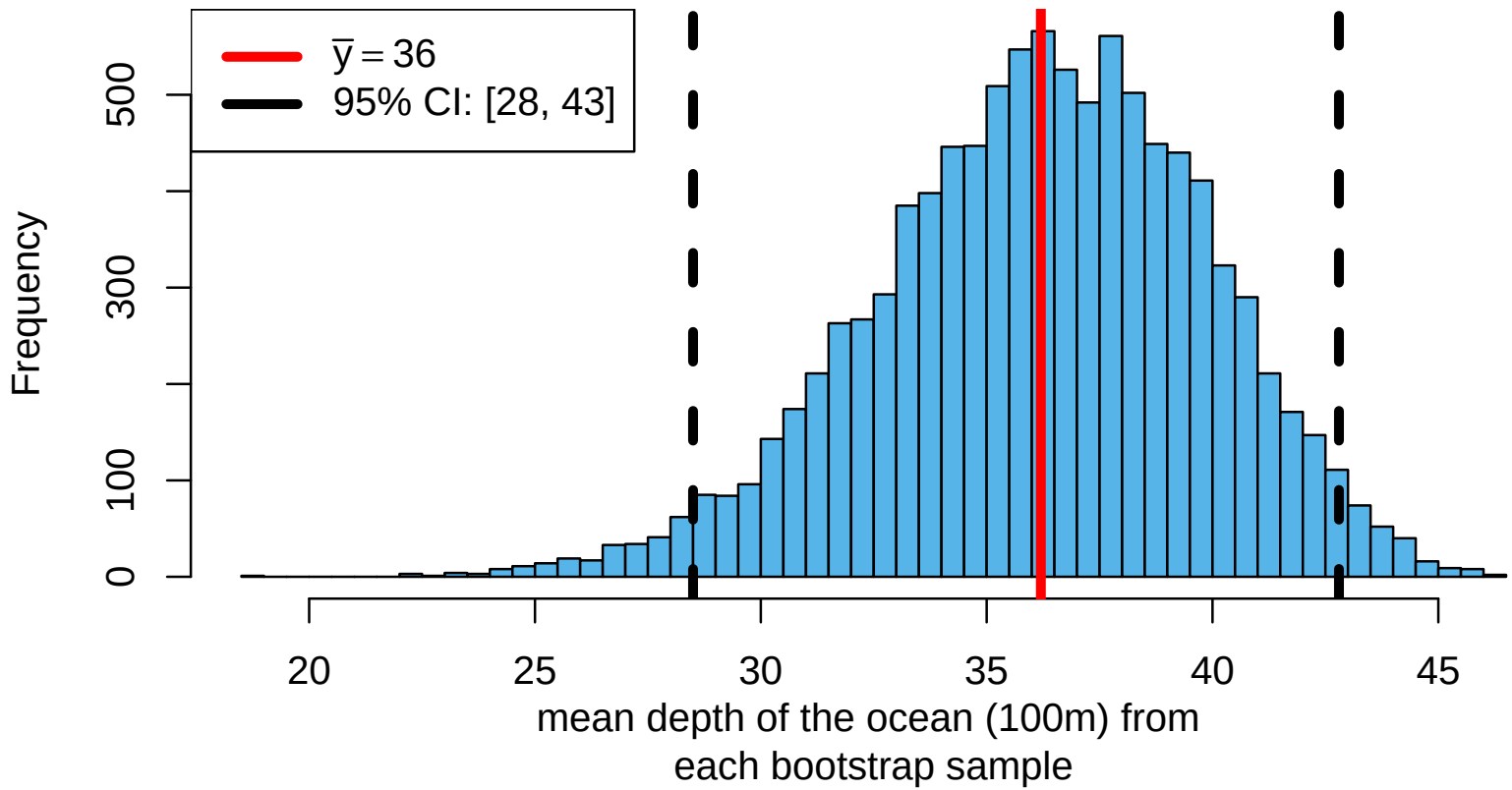


Fig.: Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{y}$), not the parameter (μ).

Main idea: simulate your own sampling distribution

```
library(mosaic)
s_dist <- do(10000) * mean( ~ alt, data = resample(depths.n.20))
CI_95 <- quantile(~ mean, data = s_dist, probs = c(0.025, 0.975))
```



One sample mean

σ known vs. unknown

σ	known	unknown
Data	$\{y_1, y_2, \ldots, y_n\}$	$\{y_1, y_2, \ldots, y_n\}$
Pop'n param	μ	μ
Estimator	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
SD	σ	$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$
SEM	$\sigma/\sqrt{n}$	$s/\sqrt{n}$
$(1-\alpha)100\%$ CI	$\bar{y} \pm z^{\star}_{1-\alpha/2}(\text{SEM})$	$\bar{y} \pm t^{\star}_{1-\alpha/2,(n-1)}(\text{SEM})$
test statistic	$\frac{\bar{y}-\mu_0}{\text{SEM}} \sim \mathcal{N}(0,1)$	$\frac{\bar{y}-\mu_0}{\text{SEM}} \sim t_{(n-1)}$

Assumptions

	z	t	Bootstrap
SRS	✓	✓	✓
Normal population	✓*	✓*	✗
needs CLT	✓*	✓*	✗
σ known	✓	✗	✗
Sampling dist. center at	μ	μ	$\bar{y}$
SD	σ	s	s
SEM	$\sigma/\sqrt{n}$	$s/\sqrt{n}$	SD(bootstrap statistics)

[a] *If population is Normal then CLT is not needed. If population is not Normal then CLT is needed.

p-values

p-values and statistical tests

Definition 3 (*p*-value)

*A **probability concerning the observed data**, calculated under a **Null Hypothesis** assumption, i.e., assuming that the only factor operating is sampling or measurement variation.*

Use To assess the evidence provided by the sample data in relation to a pre-specified claim or 'hypothesis' concerning some parameter(s) or data-generating process.

Basis As with a confidence interval, it makes use of the concept of a *distribution*.

Caution A *p*-value is NOT the probability that the null 'hypothesis' is true

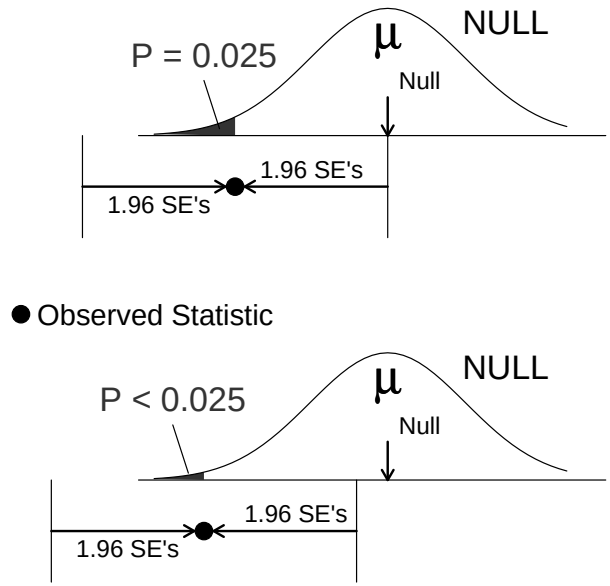
More about the *p*-value

- The *p*-value is a **probability concerning data, conditional on the Null Hypothesis being true**.
- **It is not the probability that Null Hypothesis is true,** *conditional on the data.*

$$\begin{aligned} p_{value} &= P(\text{this or more extreme data}|H_0) \\ &\neq P(H_0|\text{this or more extreme data}). \end{aligned}$$

- Statistical tests are often coded as statistically significant or not according to whether results are extreme or not with respect to a reference (null) distribution. But a test result is just one piece of data, and needs to be considered *along with rest of evidence* before coming to a 'conclusion.'
- **Likewise with statistical 'tests': the *p*-value is just one more piece of *evidence*, hardly enough to 'conclude' anything**.

Close relationship between *p*-value and CI



- (Upper graph) If upper limit of 95% CI *just touches* null value, then the 2 sided *p*-value is 0.05 (or 1 sided *p*-value is 0.025).
- (Lower graph) If upper limit *excludes* null value, then the 2 sided *p*-value is less than 0.05 (or 1 sided *p*-value is less than 0.025).
- (Graph not shown) If CI *includes* null value, then the 2-sided *p*-value is greater than (the conventional) 0.05, and thus observed statistic is "not statistically significantly different" from hypothesized null value.

Power and sample size

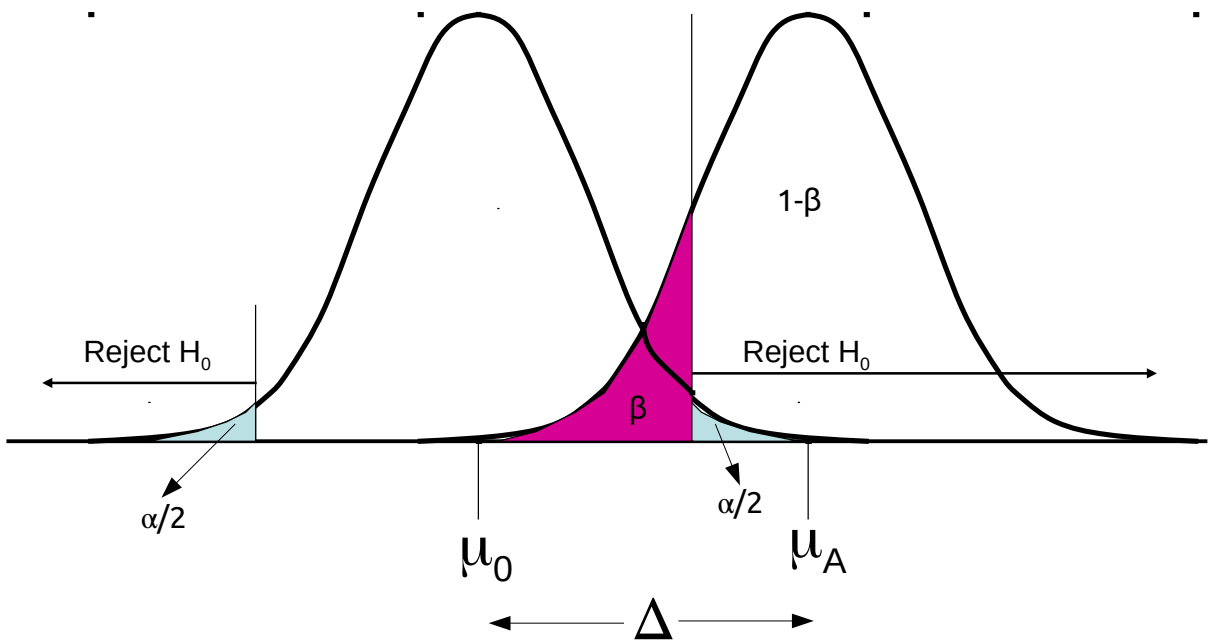
Power = $1 - \beta$

Definition 4 (Power = $1 - \beta$)

*The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect the alternative.*

Distribution of $\overline{y}$ under the null hypothesis:

Distribution of $\overline{y}$ under an alternative hypothesis:



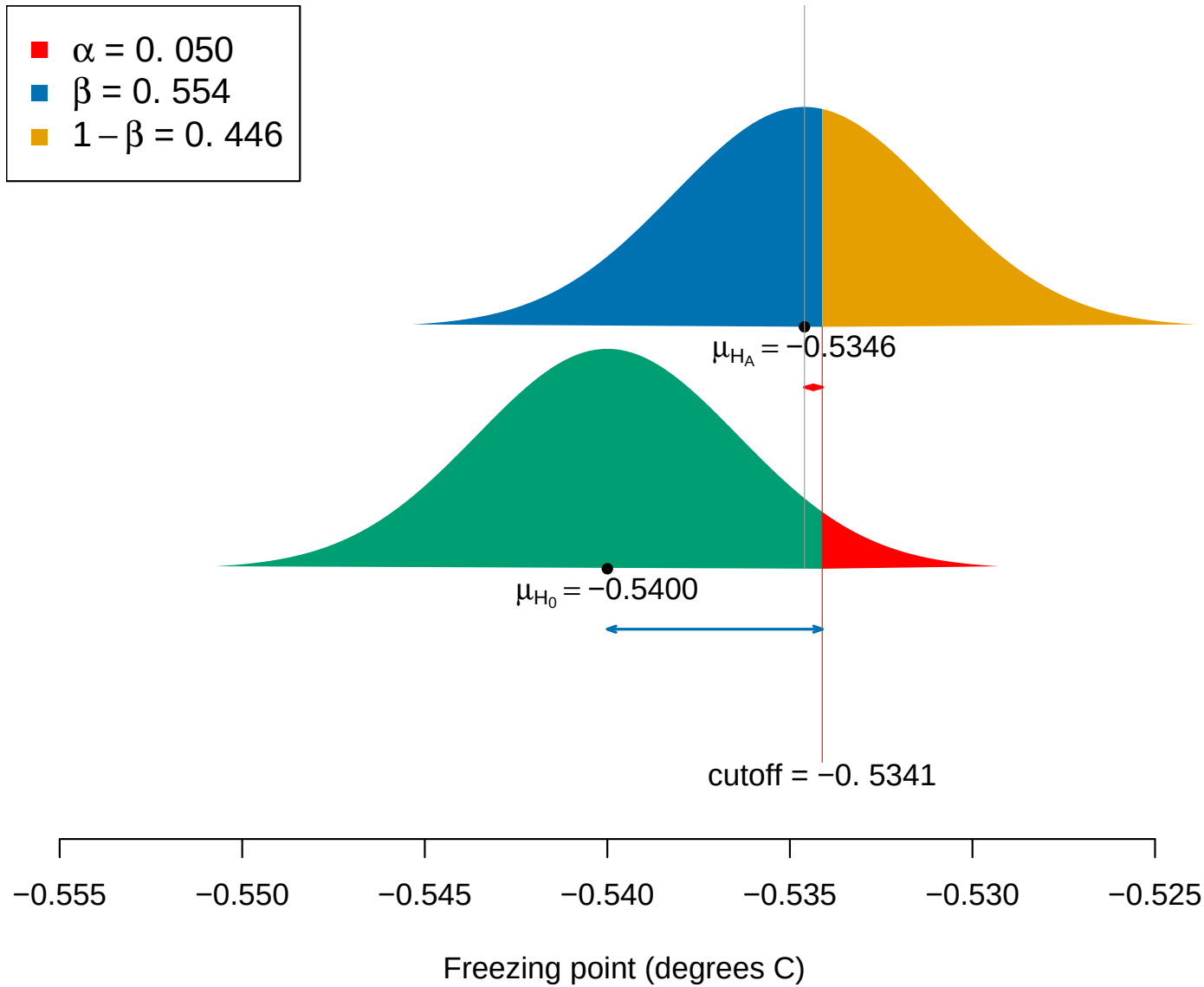
Power and Sample Size: 3 questions

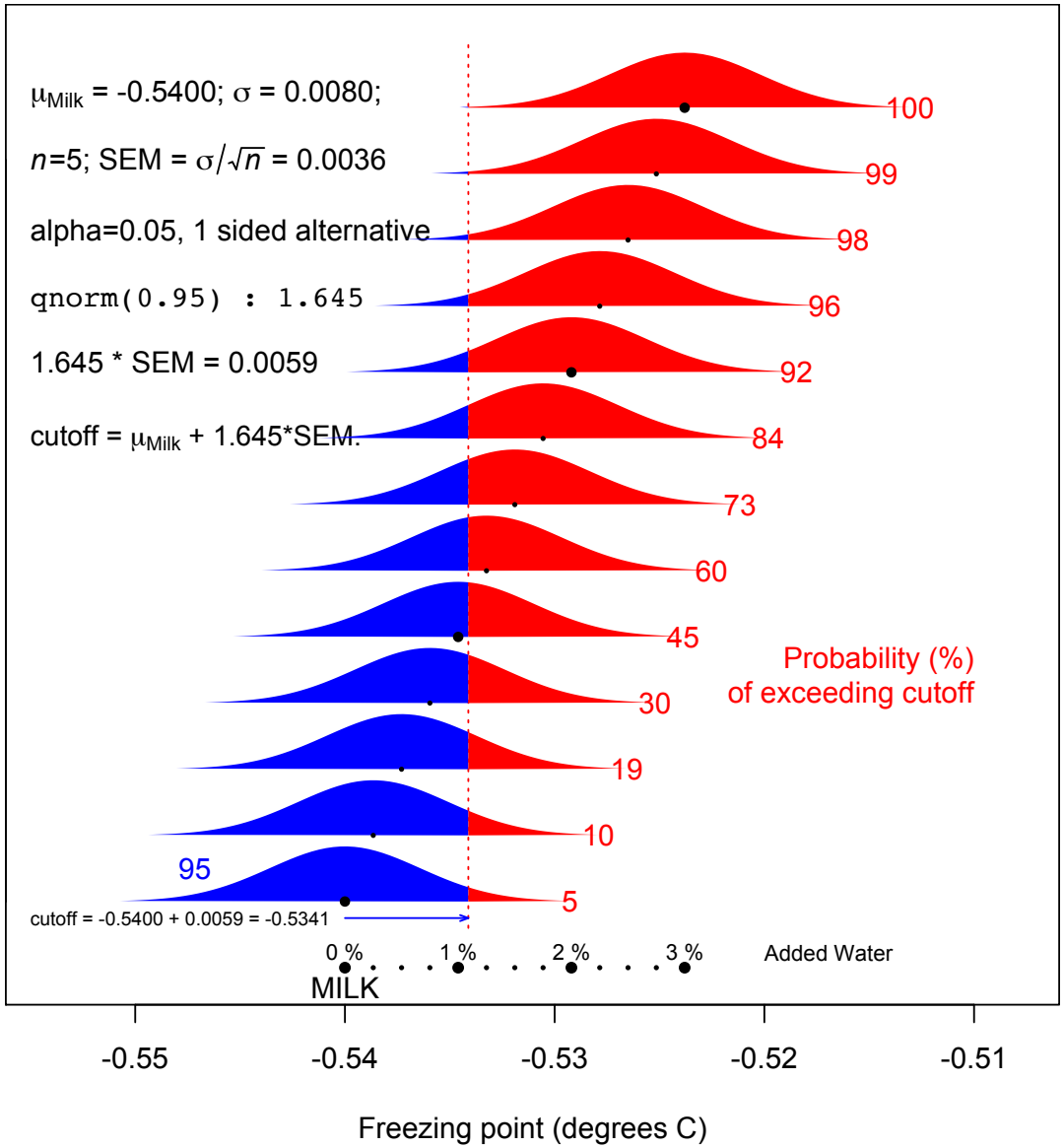
1. How much water a supplier could add to the milk before they have a 10% , 50%, 80% chance of getting caught, i.e., of the buyer detecting the cheating ?

2. Assume a 99:1 mix of milk and water. What are the chances of detecting cheating if the buyer uses samples *n*=10, 15 or 20 rather than just 5 measurements?

3. At what *n* does the chance of detecting cheating reach 80%? (*a commonly used, but arbitrary, criterion used in sample-size planning by investigators seeking funding for their proposed research*)

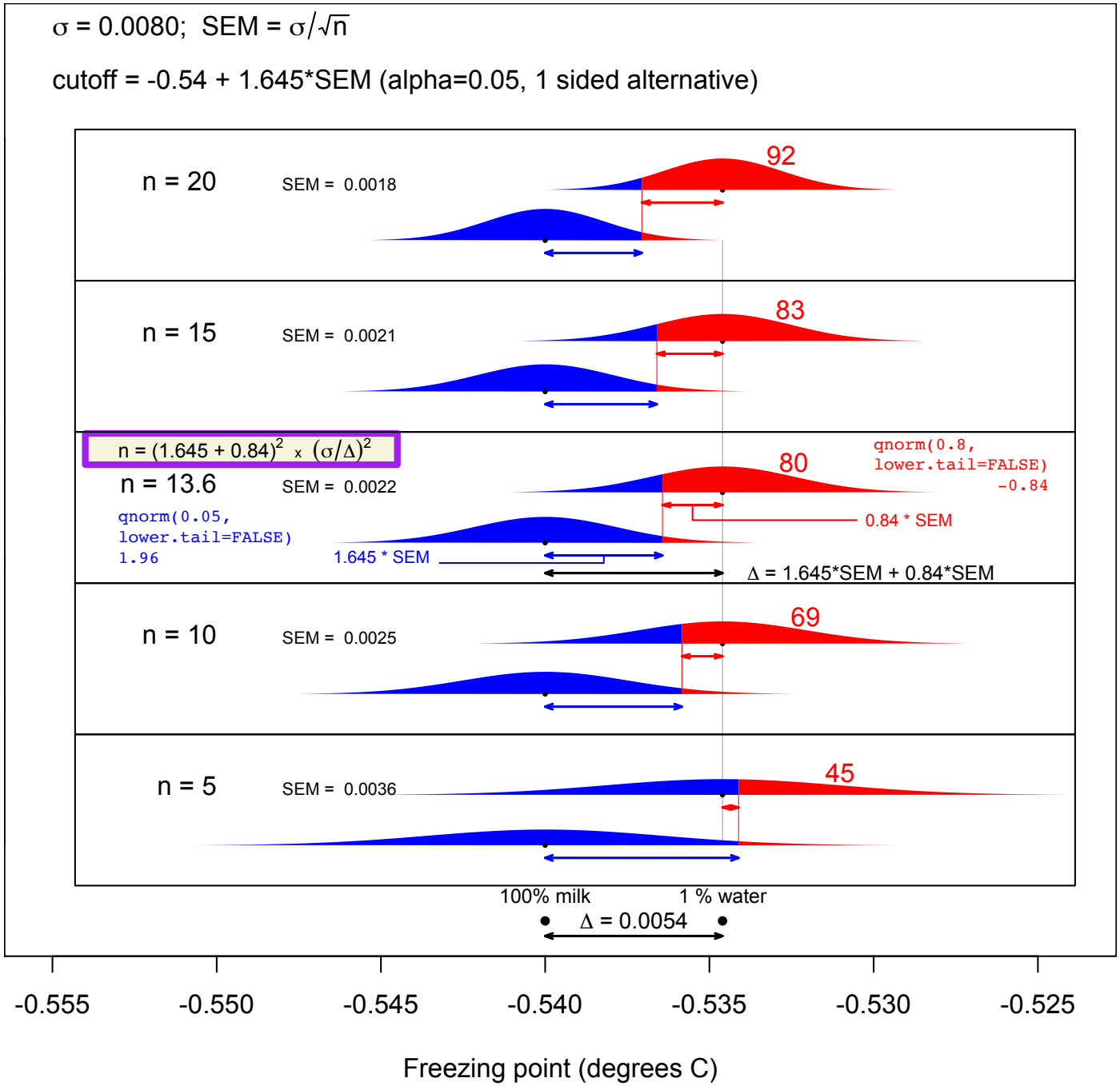
If the supplier added 1% water to the milk



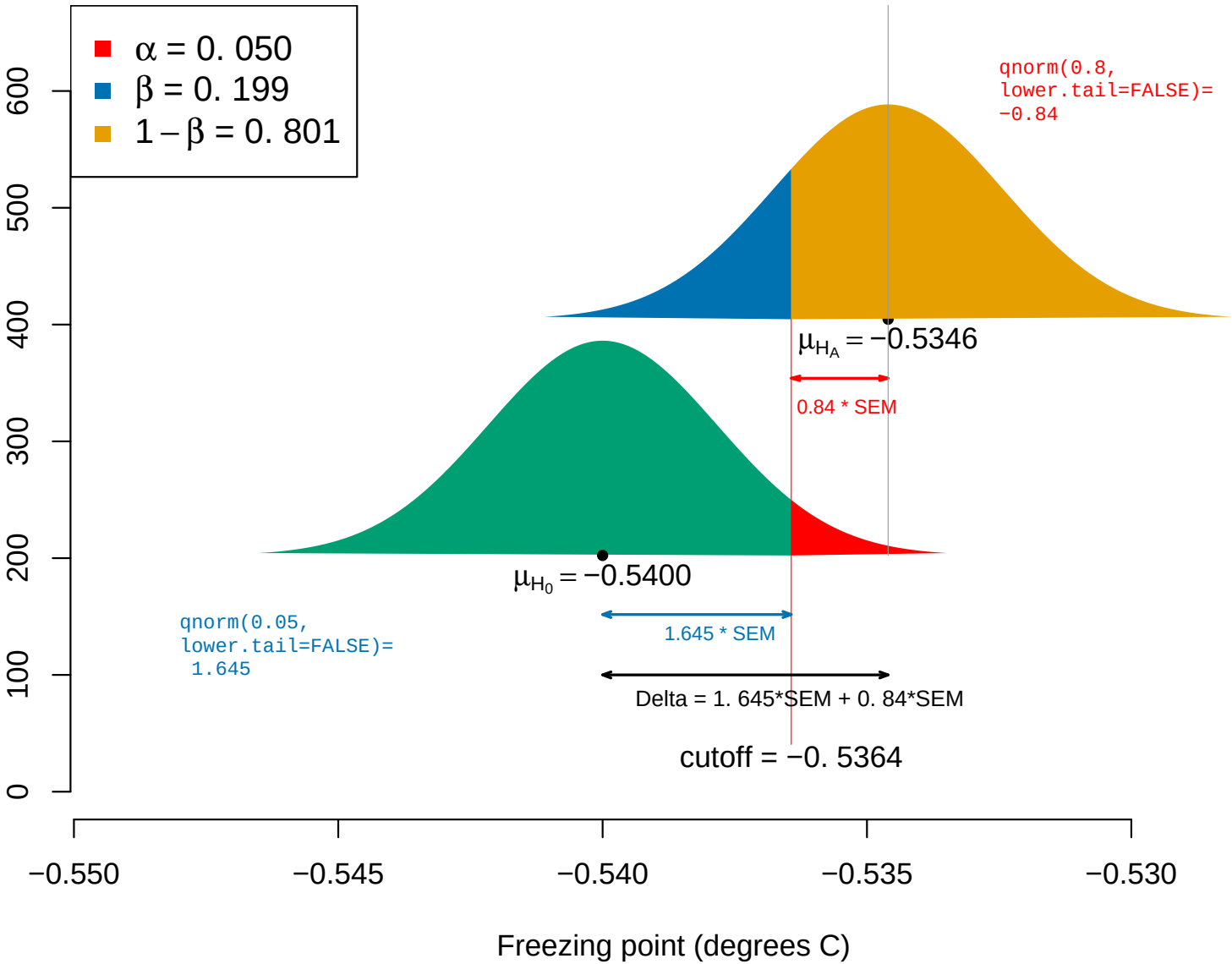


The probabilities in red were calculated using the formula:

`stats::pnorm(cutoff, mean = mu.mixture, sd = SEM, lower.tail=FALSE)`


σ = 0.0080; SEM = σ/√n
cutoff = -0.54 + 1.645*SEM (alpha=0.05, 1 sided alternative)
n = 20
SEM = 0.0018
92
n = 15
SEM = 0.0021
83
n = (1.645 + 0.84)² x (σ/Δ)²
n = 13.6
SEM = 0.0022
80
qnorm(0.8, lower.tail=FALSE)
-0.84
0.84 * SEM
qnorm(0.05, lower.tail=FALSE)
1.96
1.645 * SEM
Δ = 1.645*SEM + 0.84*SEM
n = 10
SEM = 0.0025
69
n = 5
SEM = 0.0036
45
100% milk
1 % water
Δ = 0.0054
-0.555
-0.550
-0.545
-0.540
-0.535
-0.530
-0.525
Freezing point (degrees C)

The balancing formula



What sample size needed?

- The 'balancing formula', in SEM terms, is simply the n where

$$1.645 \times SEM + 0.84 \times SEM = \Delta.$$

Replacing each of the SEMs (assumed equal, because we assumed the variability is approx. the same under both scenarios) by $\sigma/\sqrt{n}$, i.e.,

$$1.645 \times \sigma/\sqrt{n} + 0.84 \times \sigma/\sqrt{n} = \Delta.$$

and solving for n, one gets

$$n = (1.645 + 0.84)^2 \times \left\{ \frac{\sigma}{\Delta} \right\}^2 = (1.645 + 0.84)^2 \times \left\{ \frac{Noise}{Signal} \right\}^2.$$

What sample size needed? General Formula

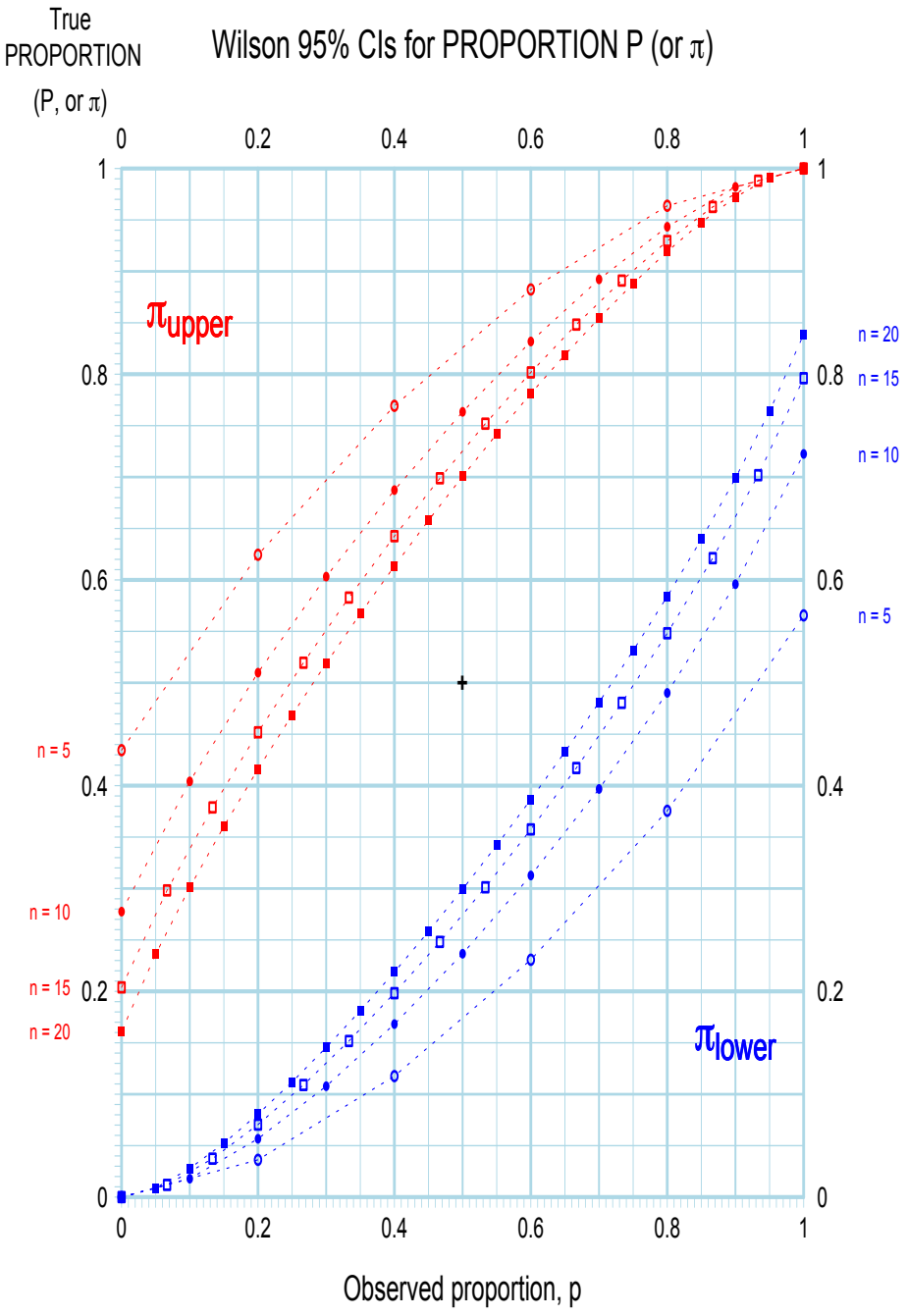
- Two sided alternative:

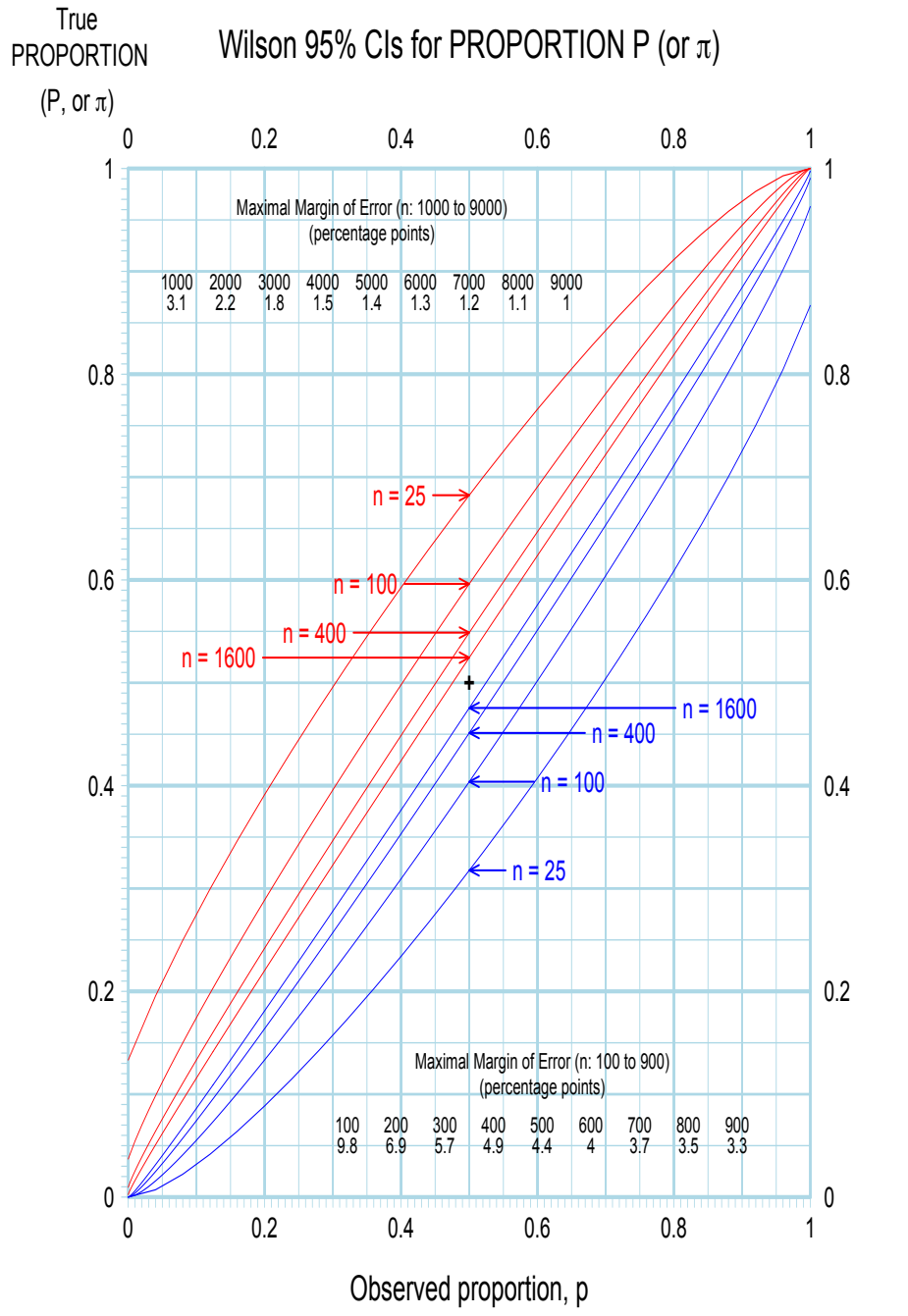
$$\Delta = z_{1-\alpha/2} \times SEM + z_{1-\beta} \times SEM$$

- One sided alternative:

$$\Delta = z_{1-\alpha} \times SEM + z_{1-\beta} \times SEM$$

One sample proportion


True
PROPORTION
(P, or π)
Wilson 95% CIs for PROPORTION P (or π)
0
0.2
0.4
0.6
0.8
1
πupper
πlower
n = 5
n = 10
n = 15
n = 20
n = 20
n = 15
n = 10
n = 5
Observed proportion, p


True
PROPORTION
(P, or π)
Wilson 95% CIs for PROPORTION P (or π)
Maximal Margin of Error (n: 1000 to 9000)
(percentage points)
1000 2000 3000 4000 5000 6000 7000 8000 9000
3.1 2.2 1.8 1.5 1.4 1.3 1.2 1.1 1
n = 25
n = 100
n = 400
n = 1600
n = 1600
n = 400
n = 100
n = 25
Maximal Margin of Error (n: 100 to 900)
(percentage points)
100 200 300 400 500 600 700 800 900
9.8 6.9 5.7 4.9 4.4 4 3.7 3.5 3.3
Observed proportion, p

Examples

Comparing two sun block lotions

Example 1

Your company produces a sun block lotion designed to protect the skin from both UVA and UVB exposure to the sun. You hire a company to compare your product with the product sold by your major competitor. The testing company exposes skin on the back of a sample of 20 people to UVA and UVB rays and measures the protection provided by each product. For 13 of the subjects, your product provided better protection. Do you have evidence to support a commercial claiming that your product provides superior UVA and UVB protection?

Comparing two sun block lotions

1. State the null hypothesis in words.

2. State the hypotheses in statistical notation
   - We need to first define the reference (null) distribution. Then the parameter of interest
   - Binomial(n=20, π=0.5) is the reference distribution where π is the proportion of people who would receive superior UVA and UVB protection from your product. The following are all equivalent:

     $H_0 : \pi = 0.5 \qquad H_a : \pi > 0.5$

     $H_0 : \pi_{\text{your product}} = \pi_{\text{their product}} = 0.5$
     $H_a : \pi_{\text{your product}} > \pi_{\text{their product}}$

     $H_0 : \pi_{\text{your product}} - \pi_{\text{their product}} = 0$
     $H_a : \pi_{\text{your product}} > \pi_{\text{their product}}$

   - You must define your own α. Here we choose $\alpha = 0.05$

Comparing two sun block lotion - p-value

1. Exact *p*-value:

```
pbinom(12, 20, 0.5, lower.tail = FALSE)
```

```
## [1] 0.131588
```

```
1 - pbinom(12, 20, 0.5)
```

```
## [1] 0.131588
```

2. Approximate *p*-value assuming Normal approximation is ok ($20 \times 0.5 \geq 10$ and $20 \times (1 - 0.5) \geq 10$)

```
SEp <- sqrt(0.5*0.5/20) # under the null
zstat <- (0.65 - 0.5) / SEp
pnorm(zstat, lower.tail = FALSE)
```

```
## [1] 0.08985625
```

Comparing two sun block lotion - Exact 95% CI

1. Exact CI (Clopper-Pearson or Nomogram):

```
mosaic::binom.test(x = 13, n = 20, p = 0.5,
ci.method = "Clopper-Pearson",
alternative = "greater")
```

```
data:  13 out of 20
number of successes = 13, number of trials = 20, p-value = 0.1316
alternative hypothesis: true probability of success is greater tha
95 percent confidence interval:
 0.4419655 1.0000000
sample estimates:
probability of success
                  0.65
```

Comparing two sun block lotion - Approximate 95% CI

1. Approximate 95% CI:

```
mosaic::binom.test(x = 13, n = 20, p = 0.5,
ci.method = "Wald",
alternative = "greater")
```

```
^^IExact binomial test (with Wald CI)

data:  13 out of 20
number of successes = 13, number of trials = 20, p-value = 0.1316
alternative hypothesis: true probability of success is greater tha
95 percent confidence interval:
 0.4745704 1.0000000
sample estimates:
probability of success
                  0.65
```

2. Approximate 95% CI assuming Normal approximation is ok

```
qnorm(c(0.025, 0.975), mean = 0.65, sd = sqrt(0.65*0.35 / 20))
```

```
## [1] 0.4409627 0.8590373
```