Final 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/

December 3, 2018

Exam Details

- **When:** Thursday December 6, 2 pm - 5 pm. Room 112 of MASS Chemistry Building
- This is a 3 hour, closed book exam.
- Two sheets of paper, double-sided, 8.5 by 11 inches are allowed.
- Calculators and translation dictionary 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.

Note on these slides

- These slides **do not** cover all the material on the final
- Review midterm review slides for material prior to midterm, and regression handouts for material after midterm

Topics to be covered

Topics to be covered

7 questions on the following topics

1. Sampling distributions, confidence intervals
2. *p*-values
3. One sample mean/rate/proportion
4. Bootstrap and CLT
5. Power
6. Standard deviation, standard error
7. Linear regression (two sample mean) → `t.test, lm` (1)
8. Poisson regression → `glm` (1)
9. Binomial regression → `glm` (1)

Standard error (SE) vs SD

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)

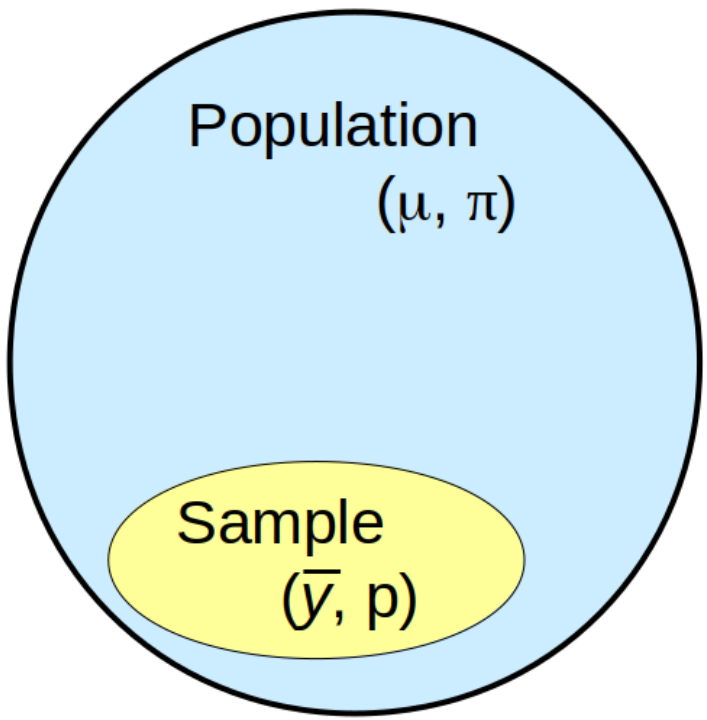
SE($\bar{y}$) describes how far $\bar{y}$ could (typically) deviate from μ;

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



Sampling Distributions

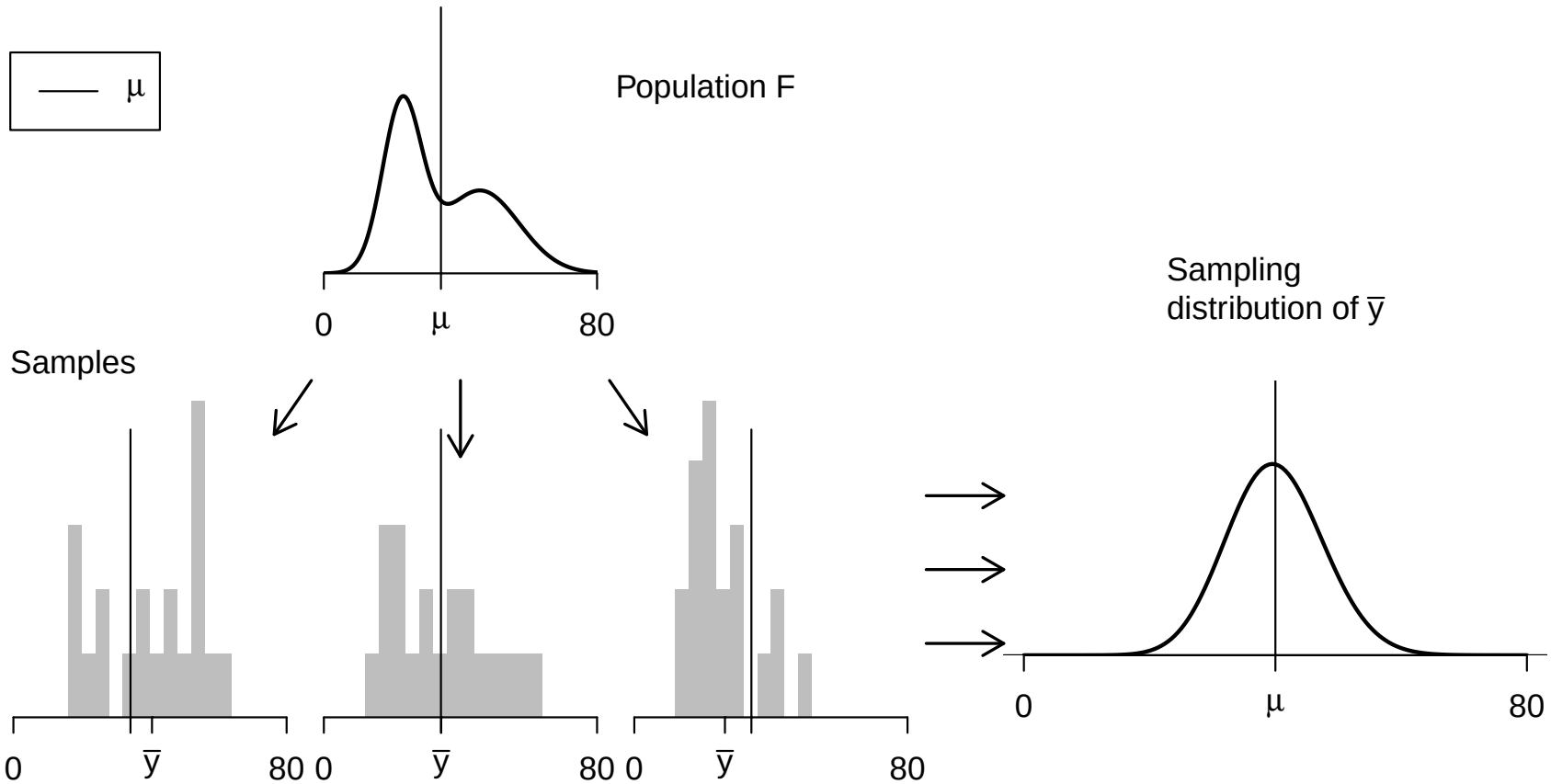


Fig.: Ideal world. Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution

Quadruple the work, half the benefit

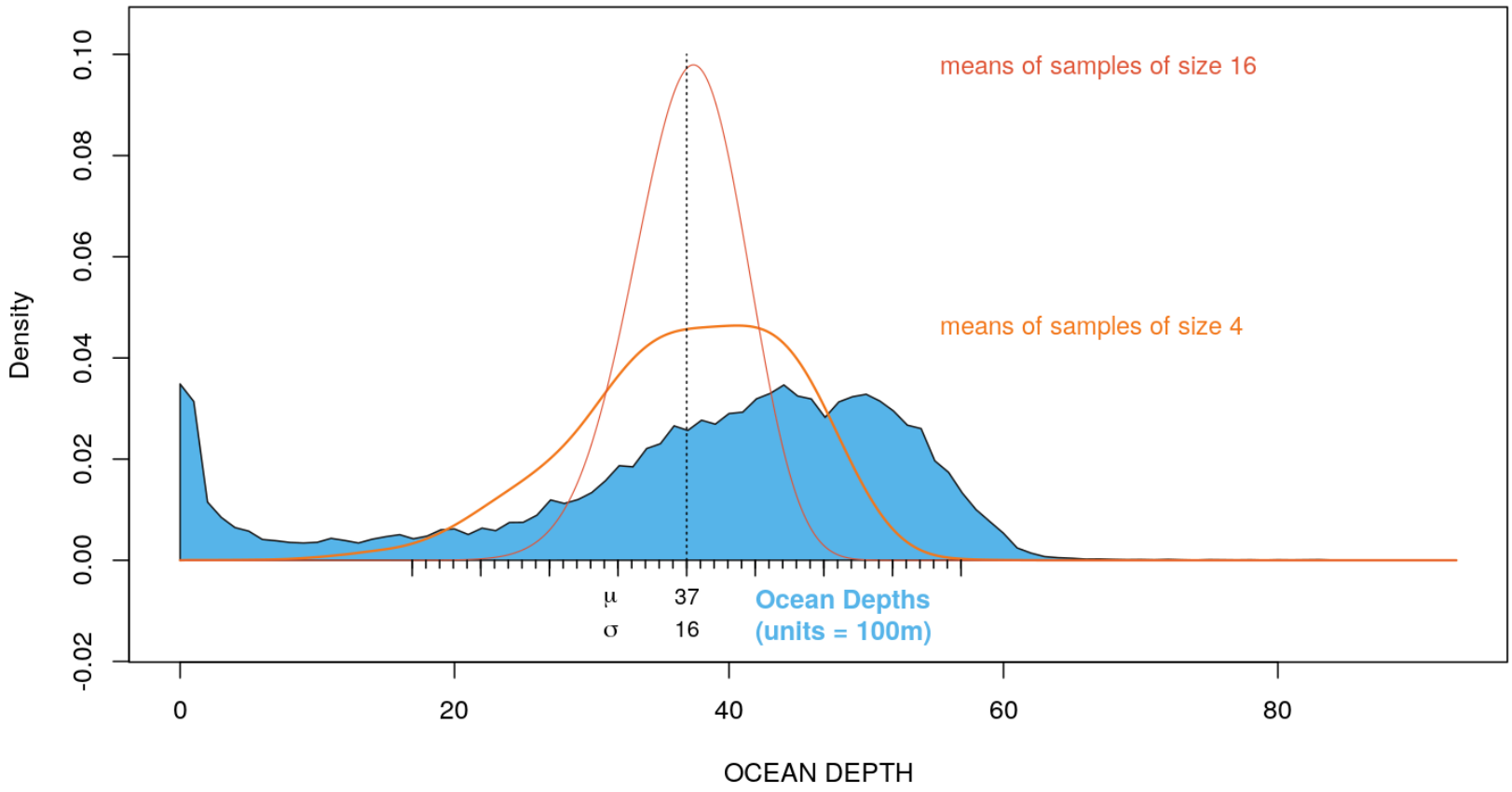


Fig.: When the sample size increases from 4 to 16, the spread of the sampling distribution for the mean is reduced by a half, i.e., the range is cut in half. This is known as the curse of the $\sqrt{n}$

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)

if $Y \sim ???(\mu_Y, \sigma_Y)$, *then*

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

Confidence Interval

Definition 1 (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

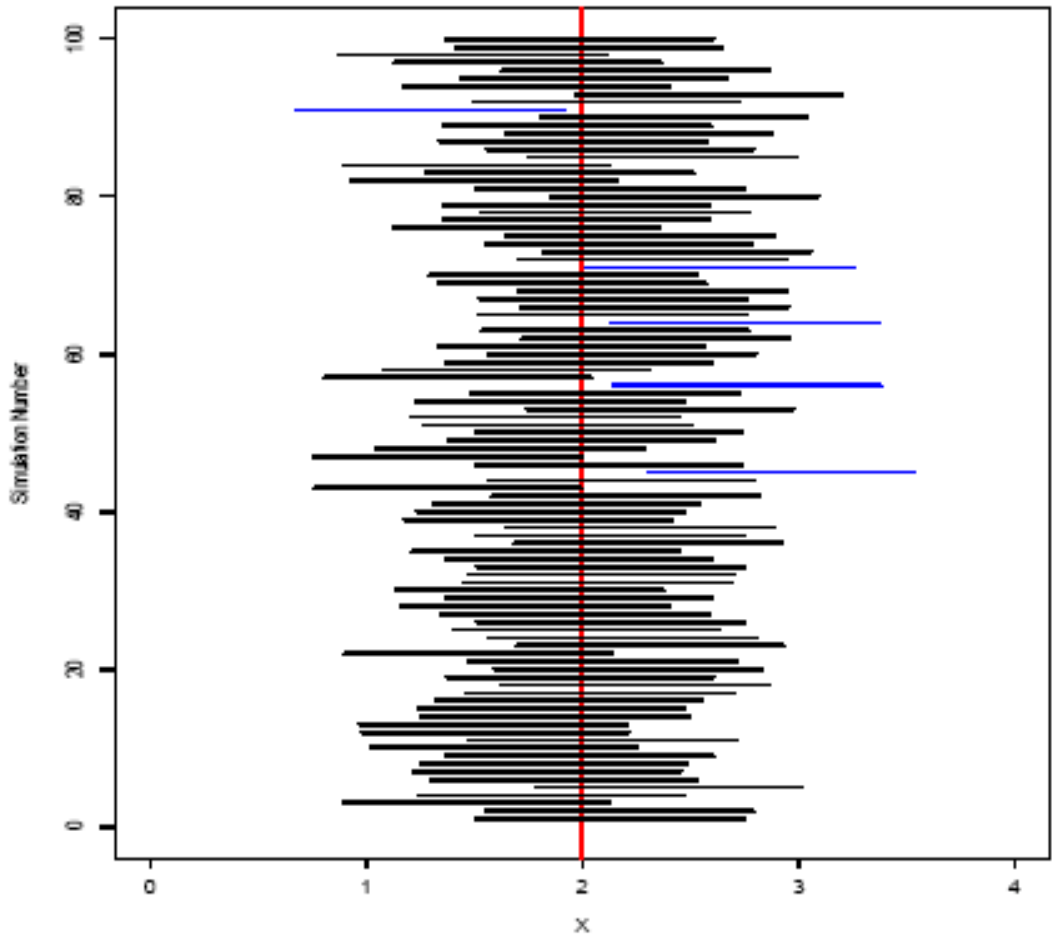


Fig.: True parameter value is 2 (red line). Each horizontal black line represents a 95% CI from a sample and contains the true parameter value. The blue CIs do not contain the true parameter value. 95% of all samples give an interval that contains the population parameter.

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`

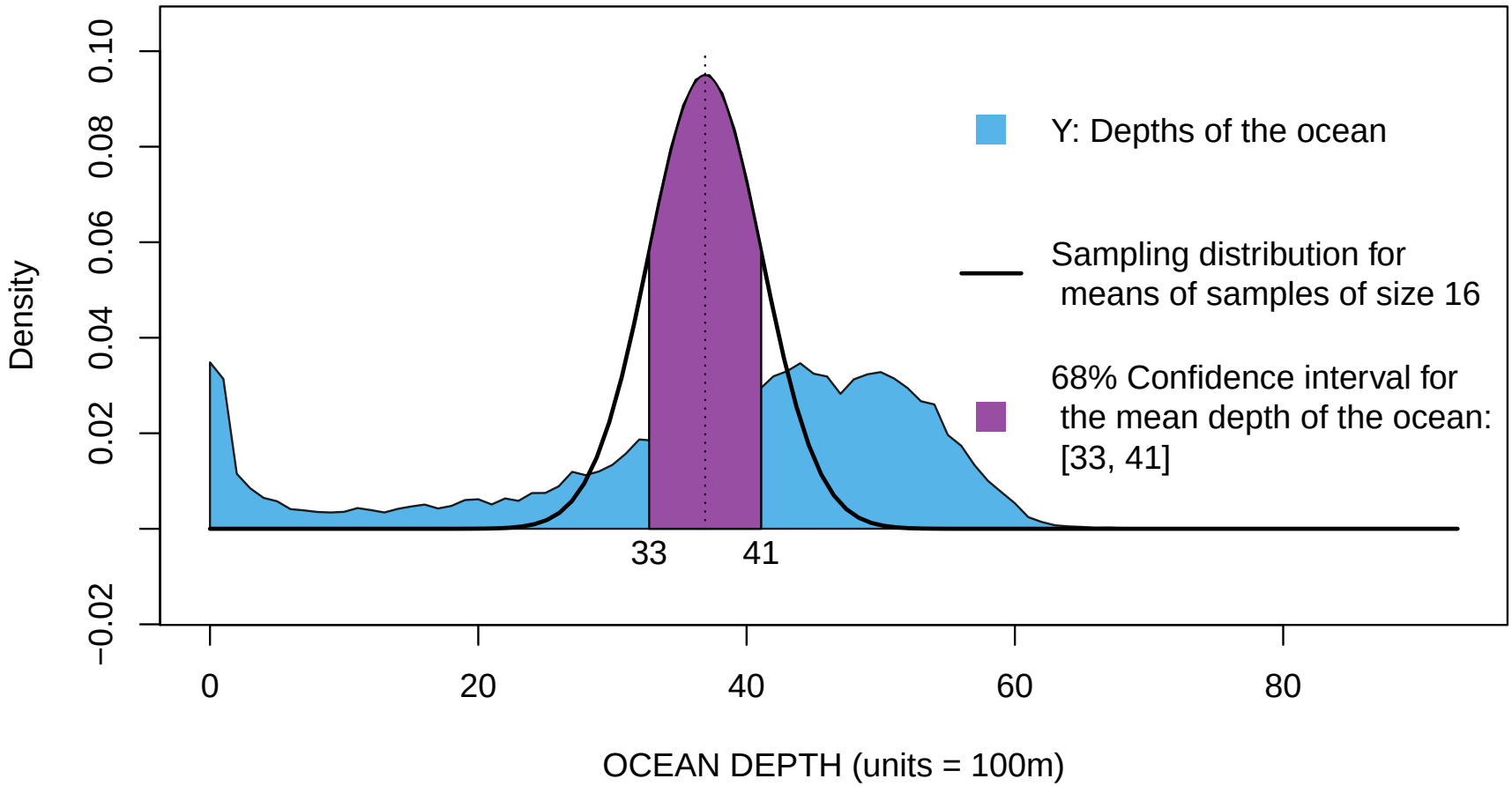


Fig.: 68% Confidence interval calculated using

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

95% Confidence interval using `qnorm`

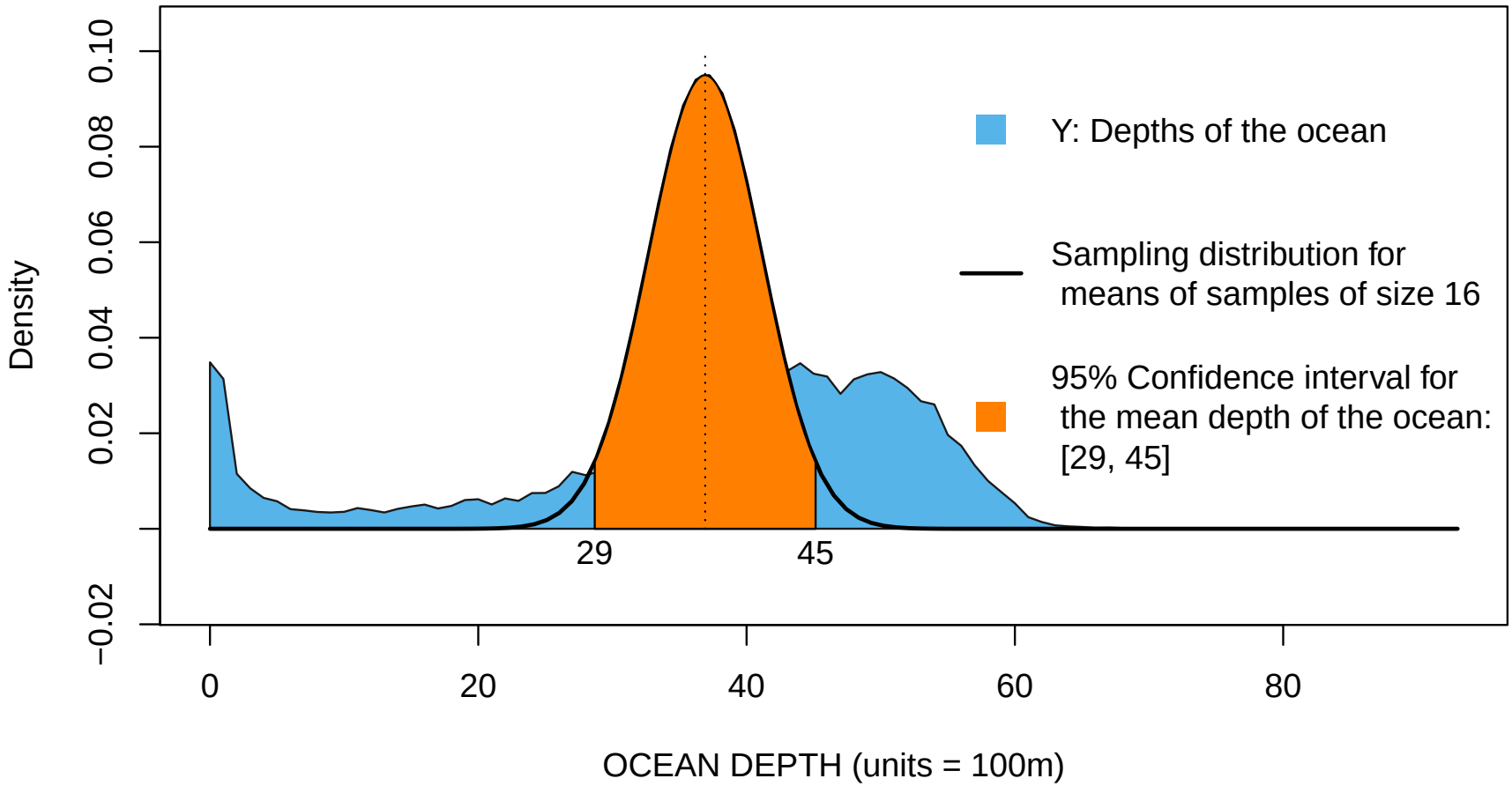


Fig.: 95% Confidence interval calculated using

```
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?

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

Reality: use the bootstrap distribution instead

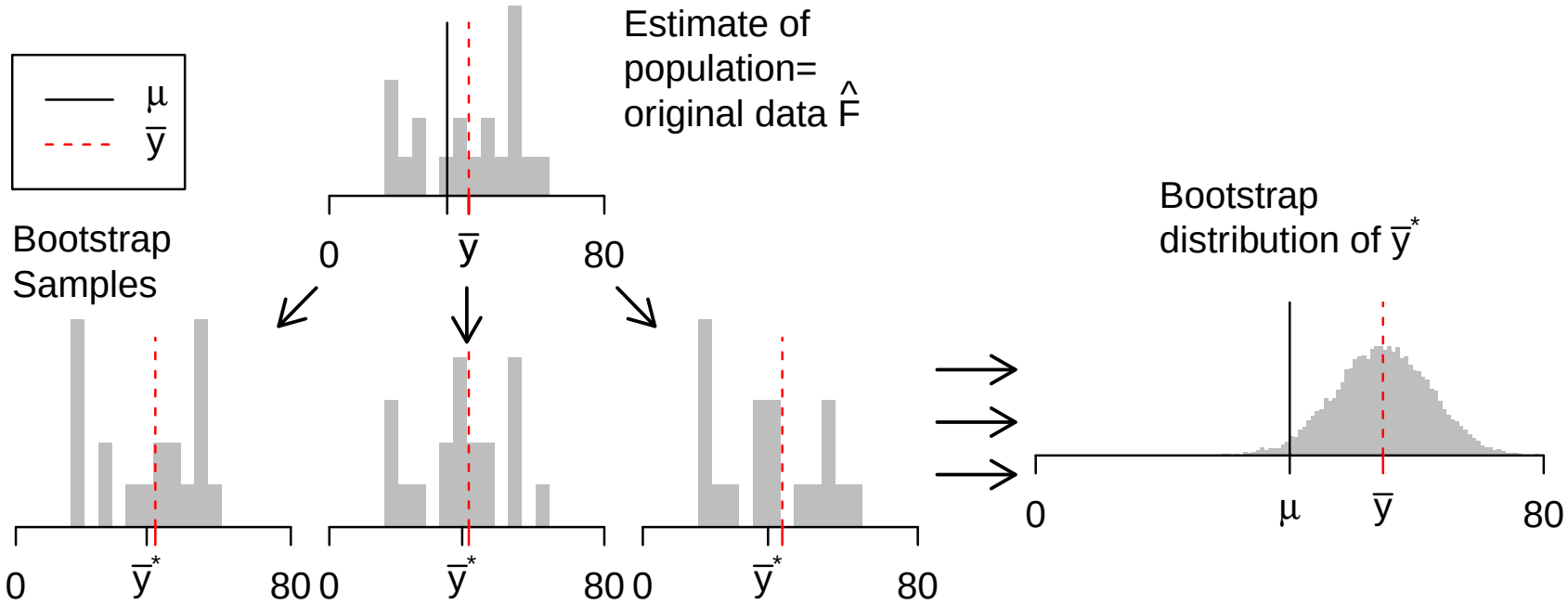
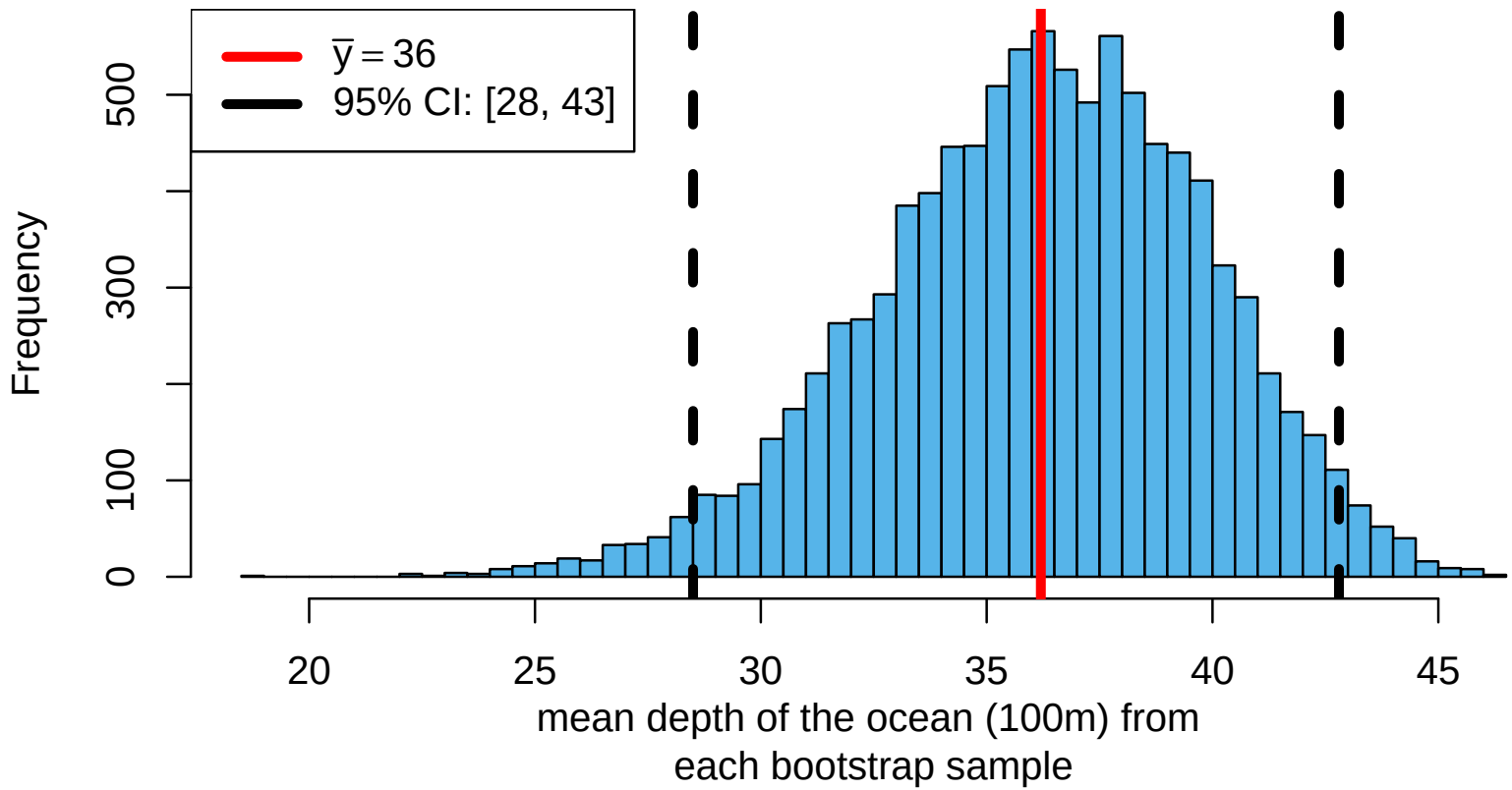


Fig.: Bootstrap world. The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{y}$), not the parameter (μ).

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))
```



Example 1: Food intake and weight gain (A5-Q1)

subject	before	after	change
1	55.7	61.7	6.0
2	54.9	58.8	3.9
3	59.6	66.0	6.4
4	62.3	66.2	3.9
5	74.2	79.0	4.8
6	75.6	82.3	6.7
7	70.7	74.3	3.6
8	53.3	59.3	6.0
9	73.3	79.1	5.8
10	63.4	66.0	2.6
11	68.1	73.4	5.3
12	73.7	76.9	3.2
13	91.7	93.1	1.4
14	55.9	63.0	7.1
15	61.7	68.2	6.5
16	57.8	60.3	2.5

```
mean(weight$change)
```

```
## [1] 4.73125
```

```
sd(weight$change) / sqrt(16)
```

```
## [1] 0.4364362
```

Example 1: Food intake and weight gain (A5-Q1)

subject	before	after	change
1	55.7	61.7	6.0
2	54.9	58.8	3.9
3	59.6	66.0	6.4
4	62.3	66.2	3.9
5	74.2	79.0	4.8
6	75.6	82.3	6.7
7	70.7	74.3	3.6
8	53.3	59.3	6.0
9	73.3	79.1	5.8
10	63.4	66.0	2.6
11	68.1	73.4	5.3
12	73.7	76.9	3.2
13	91.7	93.1	1.4
14	55.9	63.0	7.1
15	61.7	68.2	6.5
16	57.8	60.3	2.5

```
mean(weight$change)
```

```
## [1] 4.73125
```

```
sd(weight$change) / sqrt(16)
```

```
## [1] 0.4364362
```

- 95% CI for the mean weight change: $4.73 \pm$ qt(p = c(0.025, 0.975), df = 16-1)$\times$ 0.44
- p-value: pt(q = (4.73 - 0)/0.44, df=16-1, lower.tail=F)$\times 2 <$ 0.001
- This is a paired design $\rightarrow$ statistically valid to take difference and perform one-sample inference
- You can also bootstrap the **change**.

Example 1: Food intake and weight gain (A5-Q1) contd.

subject	Time	value
1	0	55.7
1	1	61.7
2	0	54.9
2	1	58.8
3	0	59.6
3	1	66.0
4	0	62.3
4	1	66.2
5	0	74.2
5	1	79.0
6	0	75.6
6	1	82.3
7	0	70.7
7	1	74.3
8	0	53.3
8	1	59.3

Will a regression on this data provide the same results?

Example 1: Food intake and weight gain (A5-Q1) contd.

subject	Time	value
1	0	55.7
1	1	61.7
2	0	54.9
2	1	58.8
3	0	59.6
3	1	66.0
4	0	62.3
4	1	66.2
5	0	74.2
5	1	79.0
6	0	75.6
6	1	82.3
7	0	70.7
7	1	74.3
8	0	53.3
8	1	59.3

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.74       2.51   26.20   <2e-16
Time            4.73       3.55    1.33     0.19

Residual standard error: 10 on 30 degrees of freedom
Multiple R-squared: 0.0559,^^IAdjusted R-squared: 0.0
F-statistic: 1.78 on 1 and 30 DF,  p-value: 0.192
```

- Point estimate is the same but standard error is much larger. Why?

Will a regression on this data provide the same results?

Example 1: Food intake and weight gain (A5-Q1) contd.

subject	Time	value
1	0	55.7
1	1	61.7
2	0	54.9
2	1	58.8
3	0	59.6
3	1	66.0
4	0	62.3
4	1	66.2
5	0	74.2
5	1	79.0
6	0	75.6
6	1	82.3
7	0	70.7
7	1	74.3
8	0	53.3
8	1	59.3

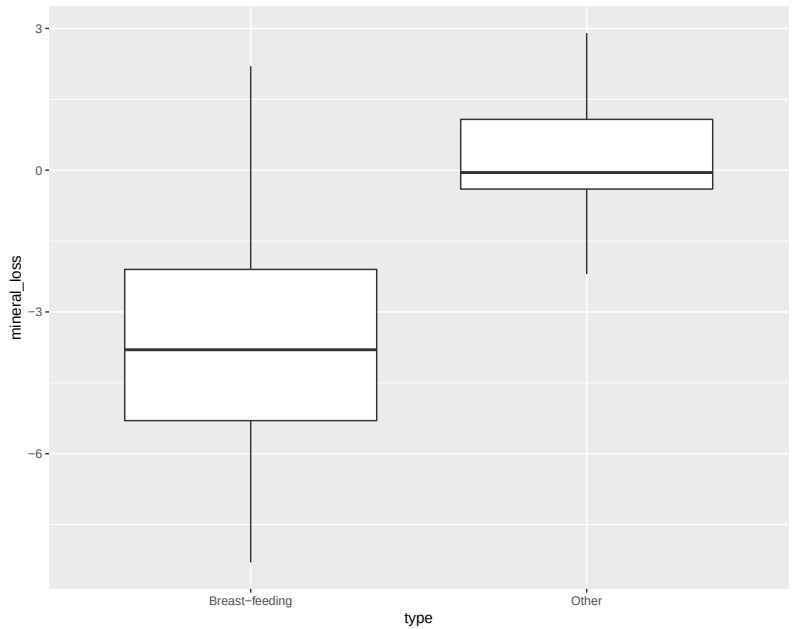
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    65.74       2.51   26.20   <2e-16
Time            4.73       3.55    1.33     0.19

Residual standard error: 10 on 30 degrees of freedom
Multiple R-squared: 0.0559,^^IAdjusted R-squared: 0.0
F-statistic: 1.78 on 1 and 30 DF,  p-value: 0.192
```

- Point estimate is the same but standard error is much larger. Why?
- Hint: Think about assumptions for inference.

Will a regression on this data provide the same results?

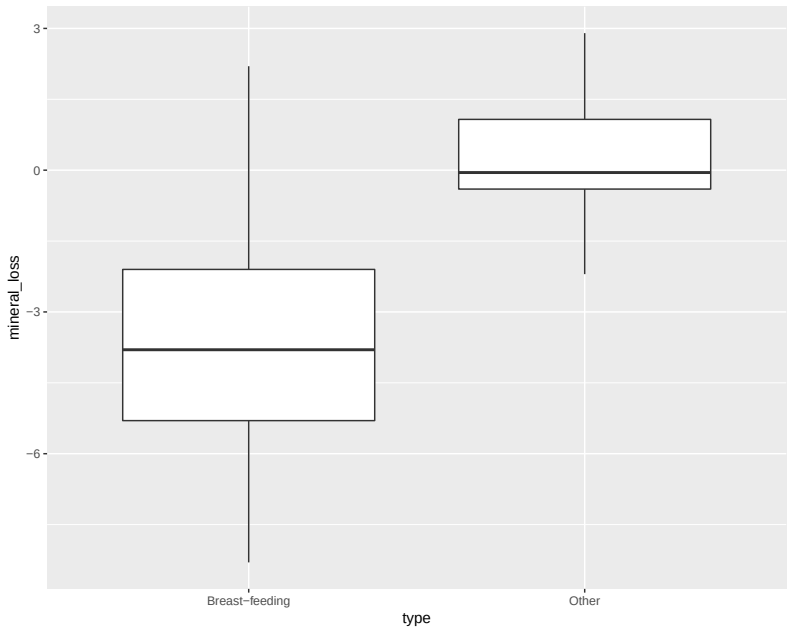
Ex. 2: Does breast-feeding weaken bones? (A4-Q3)

	type	mineral_loss
1	Other	2.4
5	Other	1.0
47	Breast-feeding	-5.2
48	Breast-feeding	-2.0
49	Breast-feeding	-2.1



Ex. 2: Does breast-feeding weaken bones? (A4-Q3)

	type	mineral_loss
1	Other	2.4
5	Other	1.0
47	Breast-feeding	-5.2
48	Breast-feeding	-2.0
49	Breast-feeding	-2.1



- Researchers compared 47 breast-feeding women with 22 women of similar age who were neither pregnant nor lactating.
- Is this a paired design?
- How can we test if the data show distinctly greater bone mineral loss among the breast-feeding women?

Ex. 2 contd. (A4-Q3)

- We could run a linear regression (equivalently a two-sample t.test with equal variances):

```
## Call: lm(formula = mineral_loss ~ type, data = boneloss)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.587      0.321  -11.18  < 2e-16
## typeOther      3.896      0.568    6.86  2.7e-09
##
## Residual standard error: 2.2 on 67 degrees of freedom
## Multiple R-squared: 0.412,^^IAdjusted R-squared: 0.404
## F-statistic:   47 on 1 and 67 DF,  p-value: 2.73e-09
```

```
# remember that var.equal=FALSE is the default in t.test
t.test(mineral_loss ~ type, data = boneloss, var.equal = TRUE)
```

```
##  Two Sample t-test with mineral_loss by type
## t = -6.8569, df = 67, p-value = 2.73e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.030524 -2.762126
## sample estimates:
## mean in group Breast-feeding          mean in group Other
##                   -3.5872340                    0.3090909
```

Ex. 2 contd. (A4-Q3)

- Two-sample t.test with unequal variances

```
# remember that var.equal=FALSE is the default in t.test
t.test(mineral_loss ~ type, data = boneloss)
```

```
## Welch Two Sample t-test with mineral_loss by type
## t = -8.4985, df = 66.197, p-value = 3.325e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.811641 -2.981008
## sample estimates:
## mean in group Breast-feeding          mean in group Other
##                   -3.5872340                    0.3090909
```

Ex. 2 contd. (A4-Q3)

- Two-sample t.test with unequal variances

```
# remember that var.equal=FALSE is the default in t.test
t.test(mineral_loss ~ type, data = boneloss)
```

```
## Welch Two Sample t-test with mineral_loss by type
## t = -8.4985, df = 66.197, p-value = 3.325e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.811641 -2.981008
## sample estimates:
## mean in group Breast-feeding          mean in group Other
##                   -3.5872340                    0.3090909
```

- Or we could bootstrap (if we suspected CLT hasn't kicked in, or non-normal population distributions) each group separately and calculate the means. Then take the difference of these means as the sampling distribution for the difference in bone mineral loss. (See A4-Q3 Bonus)

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)

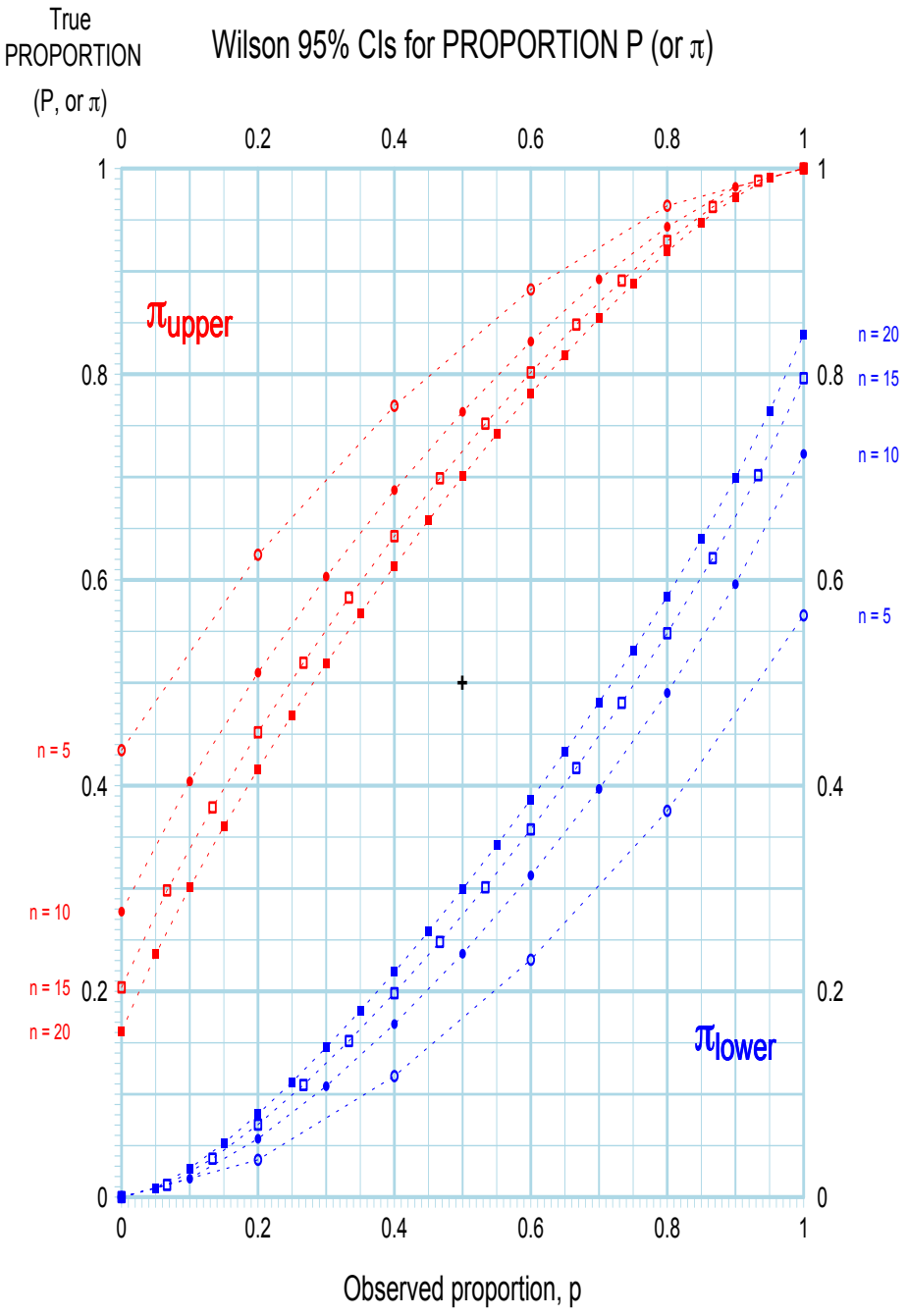
[a] *If population is Normal then CLT is not needed. If population is not Normal then CLT is needed.

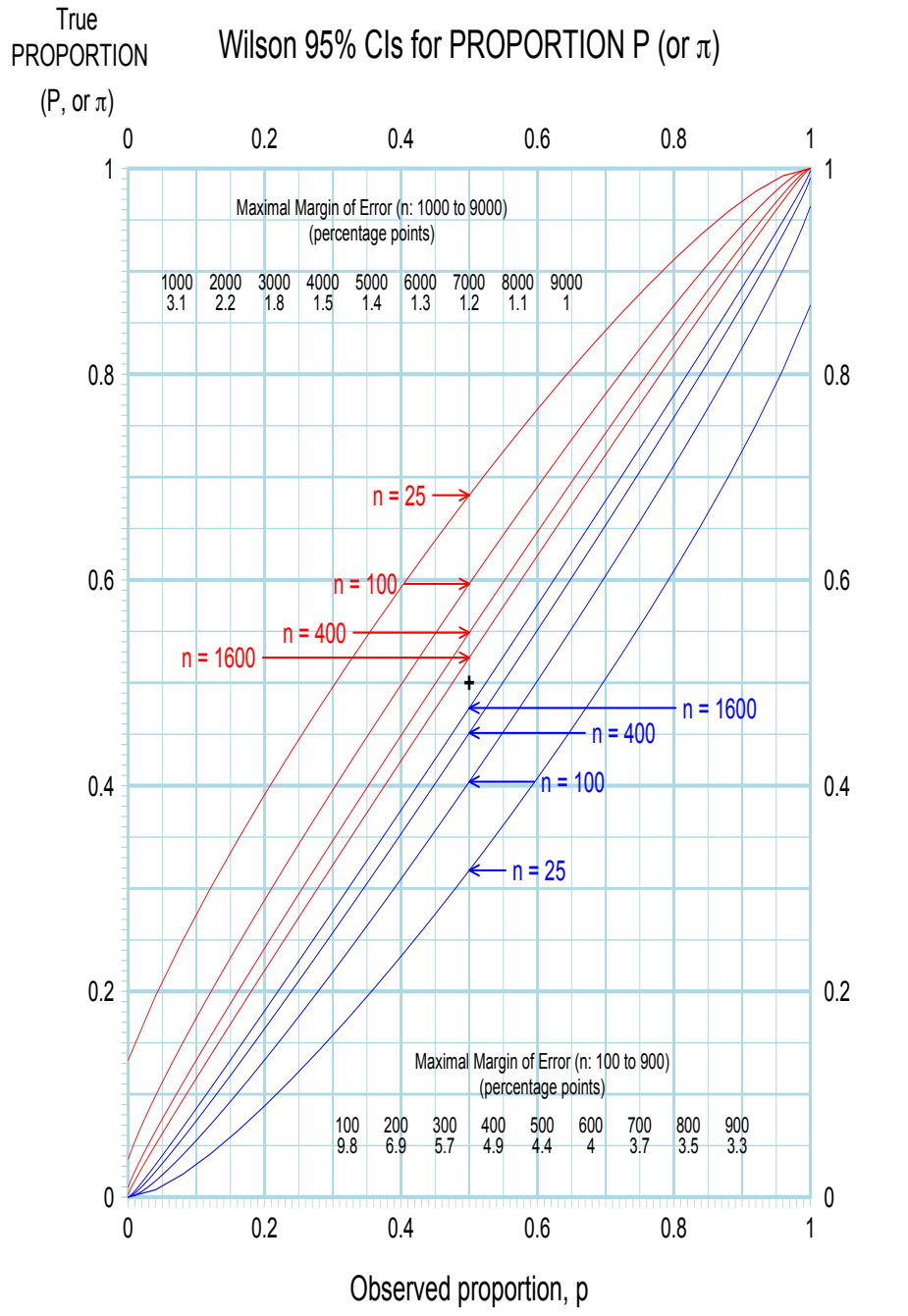
Means, Rates/Counts, Proportions

	mean	rate/count	proportion
Parameter	μ	λ/μ ($\mu = \lambda \times PT$)	π
Statistic	$\bar{y}$	$\hat{\lambda}/y$ or $\hat{\mu}$	p or $\hat{\pi}$
Distribution	Normal(μ, σ), $t_{(df)}$	Poisson(μ)	Binomial(n, π)
CI for small n	$\bar{y}\pm$ qt(·, df=n-1)×SEM	qgamma(·,shape=c(y,y+1))[c]	Clopper-Pearson
CI for large[d] n	qnorm(·, $\bar{y}$,SEM)	qnorm(·,y,$\sqrt{y}$)	qnorm(·, p, $\sqrt{\frac{p(1-p)}{n}}$)
$p$value small[a] n	pt	ppois	pbinom
$p$value large[a] n	pnorm	pnorm	pnorm

[a] need to specify lower.tail=FALSE
[b] All inference requires SRS
[c] qgamma gives CI for the count. Divide by PT if you want CI for the rate.
[d] For Normal and Poisson $n > 30$. For Binomial, $np > 10, n(1-p) > 10$

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)
0
0.2
0.4
0.6
0.8
1
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

p-values

p-values and statistical tests

Definition 2 (*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.

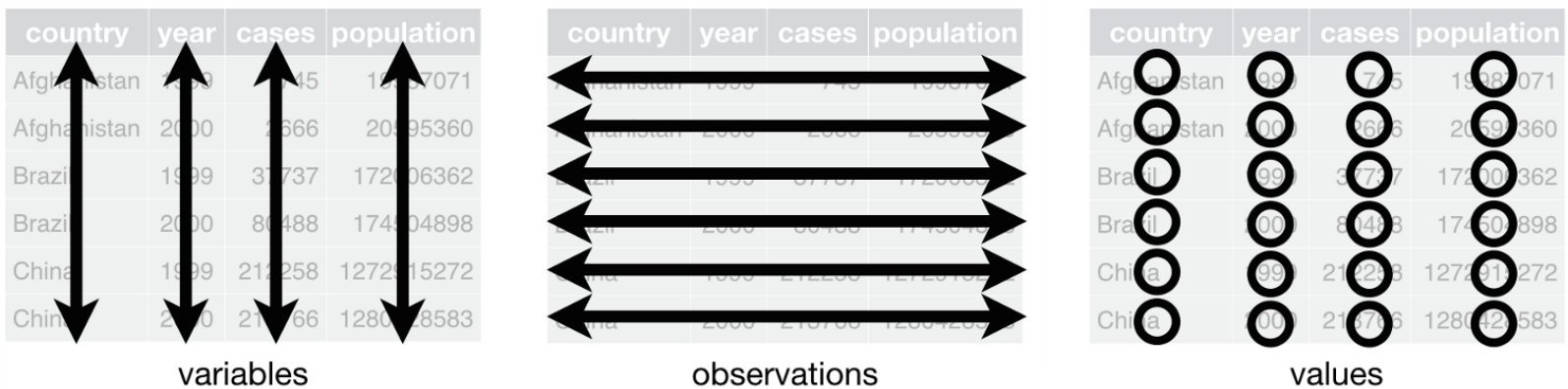
Basis — As with a confidence interval, it makes use of the concept of a *distribution*.

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

Regression

Tidy data

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational units forms a table
- Tidy data is ready for regression routines and plotting



Varia

Other remarks

- Statistical evidence $\rightarrow$ point estimate, confidence interval, p-value

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.

Comparing two sun block lotions

1. State the null hypothesis in words.

2. State the hypotheses in statistical notation

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

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.

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:

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

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:

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

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

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:

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

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

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

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
```

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")
```

```
 with 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")
```

```
Exact binomial test (with Wald CI) with 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
```