Confidence Intervals for Population Proportions

Evan L. Ray

November 1, 2017

Warm Up

Suppose $X \sim \mathrm{Normal}(\mu, \sigma)$.

Define a new random variable Z by $Z = \frac{X-\mu}{\sigma}$.

Fact: Z also follows a Normal distribution. What are the mean (i.e., expected value) and variance of Z?

"Recall" that if X is a random variable and a is a number, then

$E(aX) = aE(X)$

$E(X + a) = E(X) + a$

$\mathrm{SD}(aX) = a^2\mathrm{SD}(X)$

More Babies

- The Apgar score gives a quick sense of a baby's physical health, and is used to determine whether a baby needs immediate medical care.
- It ranges from 0 (critical health problems) to 10 (no health problems).
- Let's try to estimate the proportion of babies in the population who have an Apgar score of 10 using a sample of $n = 300$ babies.

A New Variable…

```r
babies <- mutate(babies, apgar_eq_10 = (apgar5 == 10))
head(babies[, c("gestation", "apgar5", "apgar_eq_10")])
```

```
## # A tibble: 6 x 3
##   gestation apgar5 apgar_eq_10
##       <int>  <int>       <lgl>
## 1        41      9       FALSE
## 2        47      6       FALSE
## 3        37      9       FALSE
## 4        35      9       FALSE
## 5        37     10        TRUE
## 6        35      9       FALSE
```

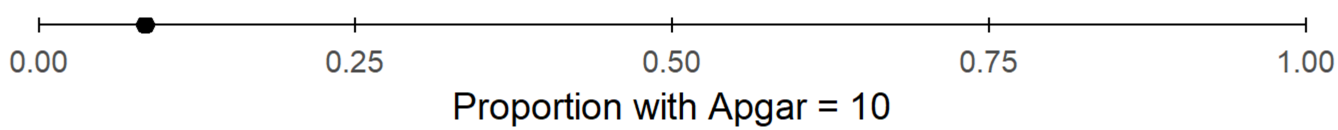
Population Proportion

```
table(babies$apgar_eq_10)
```

```
##
##  FALSE   TRUE
## 236381  21648
```

```
table(babies$apgar_eq_10) / nrow(babies)
```

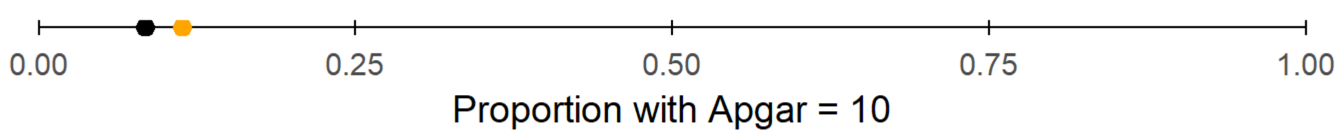
```
##
##      FALSE       TRUE
## 0.91610245 0.08389755
```



Sample Proportion

```
babies_sample <- sample_n(babies, size = 300)
table(babies_sample$apgar_eq_10) / nrow(babies_sample)
```

```
## 
##     FALSE      TRUE 
## 0.8866667 0.1133333
```



- Our estimate of the population proportion based on this sample is WRONG!
- Can we get a sense of how wrong it might be, using only the data in our sample?

Sampling Distribution of $\hat{p}$

- On Monday we said that if n is big enough,

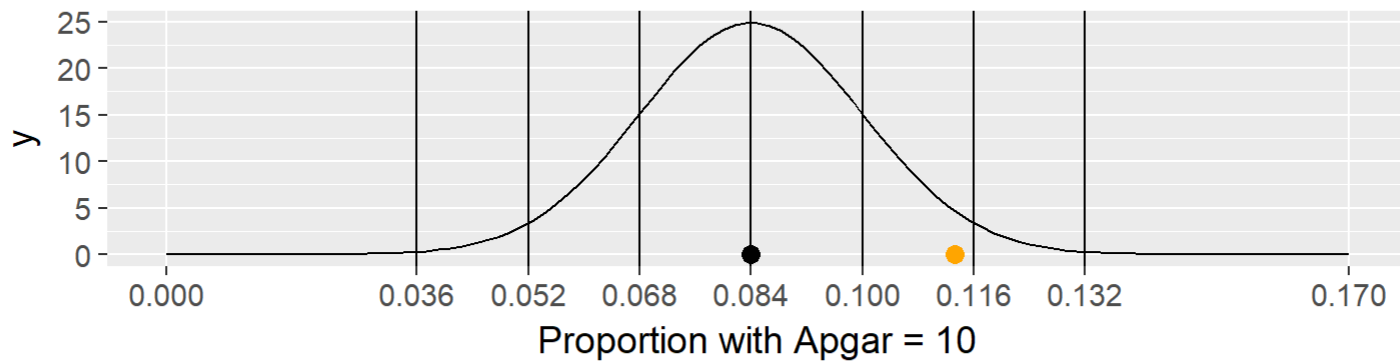
$$\hat{p} \sim \text{Normal}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- In this case, the population proportion is $p = 0.084$, and $n = 300$, so…

$$\hat{p} \sim \text{Normal}\left(0.084, 0.016\right)$$

Interpretation with 68-95-99.7 Rule

$$\hat{p} \sim \text{Normal}\,(0.084, 0.016)$$



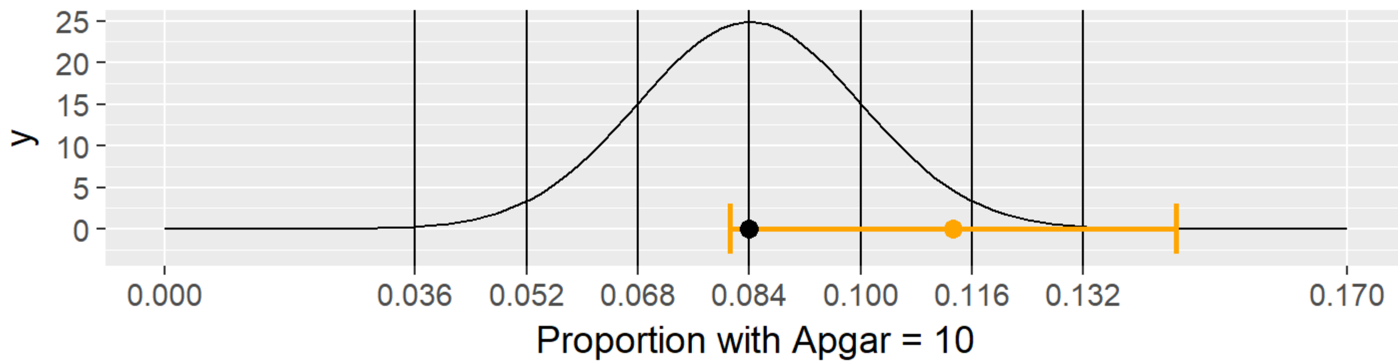
- For about 68% of samples of size n we could take, the sample proportion $\hat{p}$ will be within $\pm$ 1 standard deviation ($\pm$ 0.016) of the population proportion $p = 0.084$

- For about 95% of samples of size n we could take, the sample proportion $\hat{p}$ will be within $\pm$ 2 standard deviations ($\pm$ 0.032) of the population proportion $p = 0.084$

A Confidence Interval

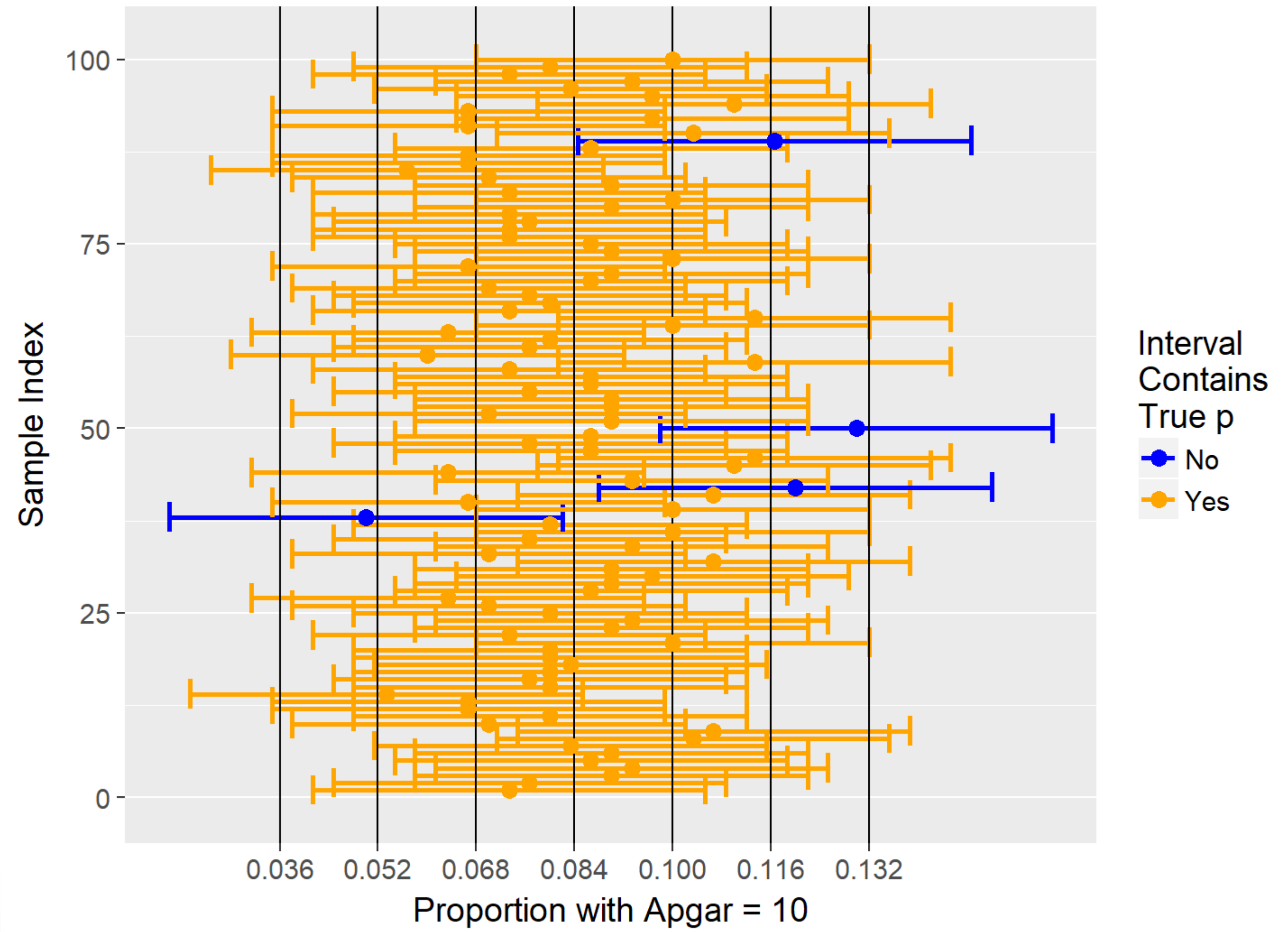
- If $\hat{p}$ is within $\pm$ 2 standard deviations of p, then p is contained in the interval

$$[\hat{p} - 2\,\mathrm{SD}(\hat{p}), \hat{p} + 2\,\mathrm{SD}(\hat{p})]$$



- We are "95% Confident" that the population proportion p is in the interval [0.081, 0.145].
- For 95% of samples, an interval constructed this way contains p.

95% C.I.s from 100 Different Samples



A Minor Problem

- The 95% confidence interval from a couple of slides ago was

$$[\hat{p} - 2\,\mathrm{SD}(\hat{p}), \hat{p} + 2\,\mathrm{SD}(\hat{p})]$$

- But $\mathrm{SD}(\hat{p})$ depends on the (unknown) population parameter p:

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

A Minor Problem

- The 95% confidence interval from a couple of slides ago was

$$[\hat{p} - 2\,\mathrm{SD}(\hat{p}), \hat{p} + 2\,\mathrm{SD}(\hat{p})]$$

- But $\mathrm{SD}(\hat{p})$ depends on the (unknown) population parameter p:

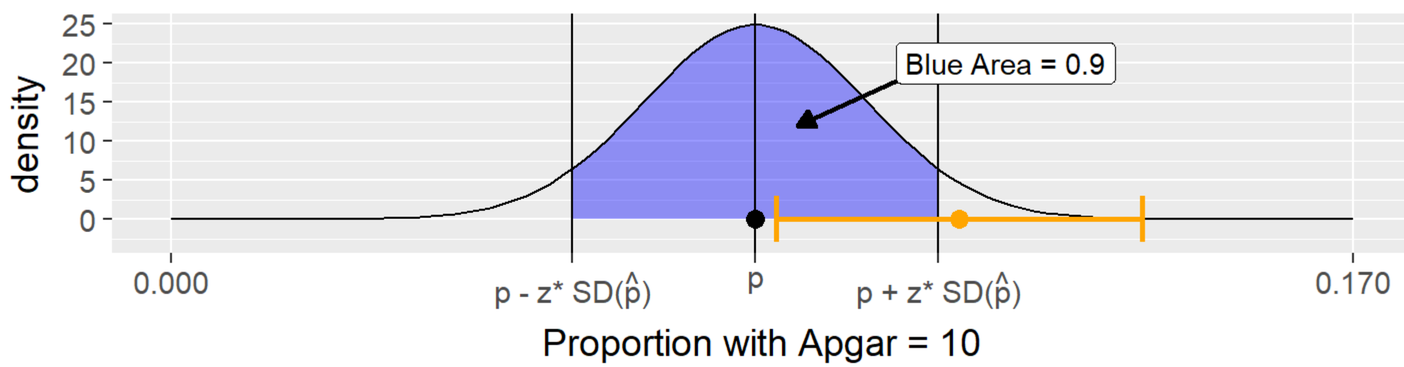
$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

- We can **estimate** $\mathrm{SD}(\hat{p})$ by plugging our **estimate** of p into this formula. An estimate of the standard deviation of a sampling distribution is called a **standard error**:

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Critical Values

- What if we want a 90% CI instead of a 95% CI?
- We need to know: 90% of sample means will be within how many standard deviations of the population mean?
- This is called the **critical value**, and denoted by z^*



- Our new CI formula: $[\hat{p} - z^* \mathrm{SE}(\hat{p}), \hat{p} + z^* \mathrm{SE}(\hat{p})]$

Finding the Critical Value (Short Version)

- For a 90% CI, the critical value is the 95th percentile of a Normal(0, 1) distribution:

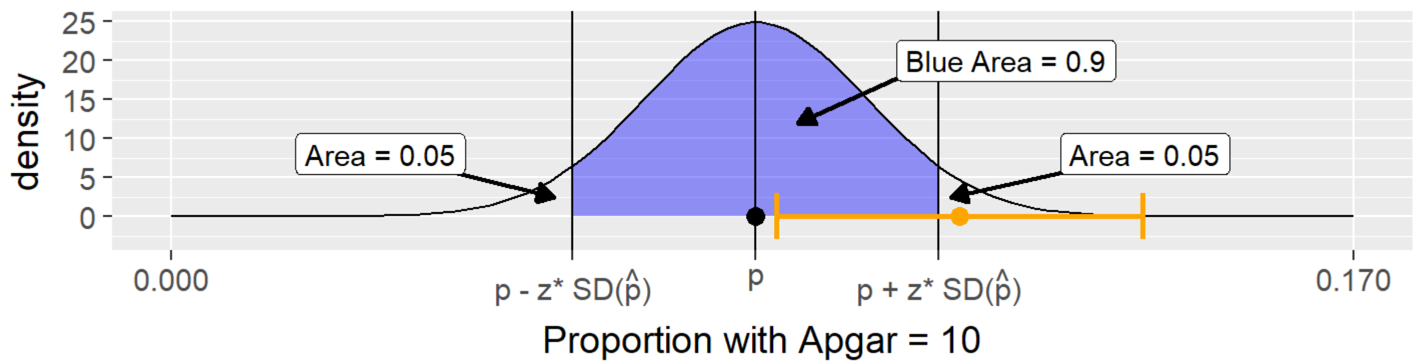
```
qnorm(0.95, mean = 0, sd = 1)
```

```
## [1] 1.644854
```

- More generally: for a $(1-\alpha) \times 100\%$ CI, the critical value is the (1 -)th quantile of a Normal(0, 1) distribution:
  - $\alpha = 0.1$ -> 90% CI. 1 - 0.05 = 0.95th quantile.
  - $\alpha = 0.05$ -> 95% CI. 1 - 0.025 = 0.975th quantile.
  - $\alpha = 0.01$ -> 99% CI. 1 - 0.005 = 0.995th quantile.

Finding the Critical Value

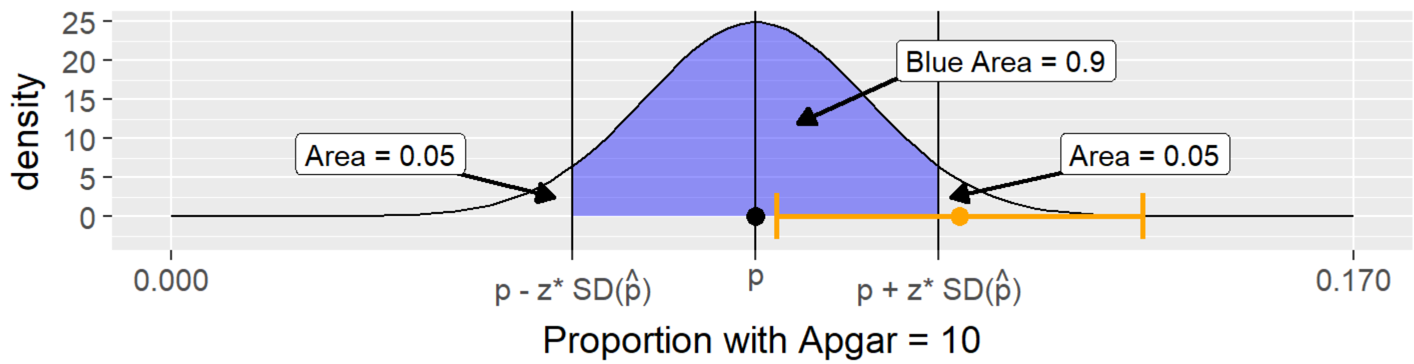
- $\hat{p} \sim \text{Normal}(p, \text{SD}(\hat{p}))$



- For a 90% CI, we need the total area to the left of $p + z^*\text{SD}(\hat{p})$ to be 0.95, in a Normal(p, SD($\hat{p}$)) distribution.

Finding the Critical Value (continued)

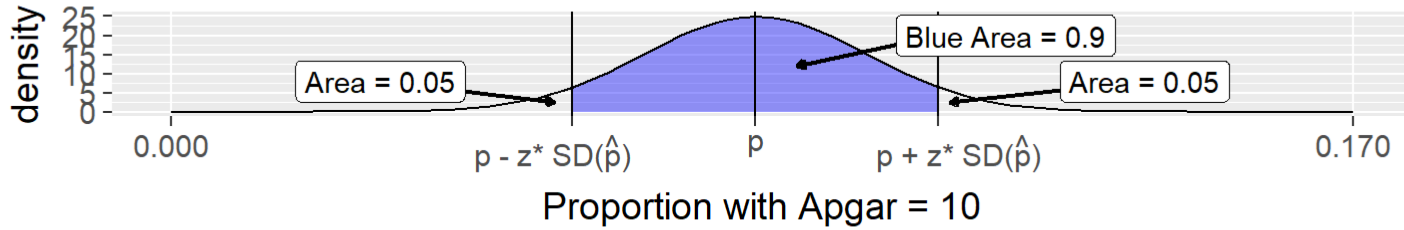
- $\hat{p} \sim \text{Normal}(p, \text{SD}(\hat{p}))$



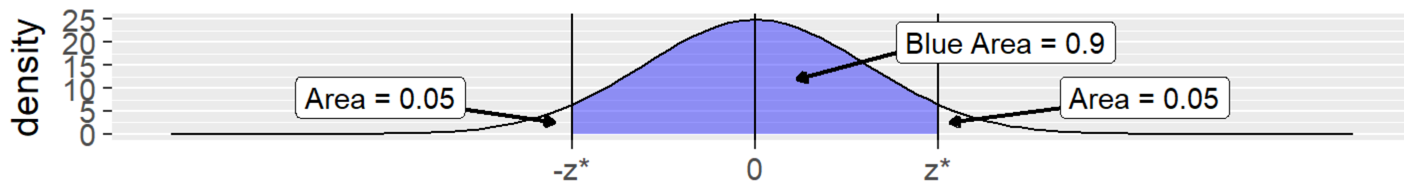
- For a 90% CI, we need the total area to the left of $p + z^*\text{SD}(\hat{p})$ to be 0.95, in a Normal(p, SD($\hat{p}$)) distribution.

- Let's define $Z = \frac{\hat{p} - p}{\text{SD}(\hat{p})}$. Then $Z \sim \text{Normal}(0, 1)$ (see warmup)

Finding the Critical Value (continued)



- For a 90% CI, area to the left of $p + z^* \mathrm{SD}(\hat{p})$ is 0.95.
- Define $Z = \frac{\hat{p} - p}{\mathrm{SD}(\hat{p})}$. Then $Z \sim \mathrm{Normal}(0, 1)$



- Area to the left of $\frac{[p + z^* \mathrm{SD}(\hat{p})] - p}{\mathrm{SD}(\hat{p})} = z^*$ is 0.95.

Putting it All Together

- **CI formula**: $[\hat{p} - z^* \mathrm{SE}(\hat{p}), \hat{p} + z^* \mathrm{SE}(\hat{p})]$
- **Standard Error** of $\hat{p}$: $\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Critical Value**: z^* is the 97.5th percentile of a standard normal distribution if we want a 95% CI
  - Use `qnorm` function in R
- **Margin of Error**: $z^* \mathrm{SE}(\hat{p})$ (how much we add and subtract from the point estimate $\hat{p}$)
- **Interpretation**: In repeated sampling, a confidence interval constructed using this procedure contains the population parameter for 95% of samples (or whatever your confidence level is).

Assumptions to Check

- Two outcomes (that are relevant to this analysis)
- Same probability of success
- People/items in our sample are **independent**
  - Think about how data were collected/if there is a connection between units
  - 10% Condition: Sample size less than 10% of population size?
- **Sample size** large enough to use normal approximation to the sampling distribution:
  - $np \geq 10$ and $n(1-p) \geq 10$
  - … but we don't actually know p!
  - Check that there are at least 10 "successes" and 10 "failures" in the data set.

Manual Calculations in R

```
table(babies_sample$apgar_eq_10) / nrow(babies_sample)
```

```
##
##     FALSE      TRUE
## 0.8866667 0.1133333
```

```
p_hat <- 0.1133333
se_p_hat <- sqrt(p_hat * (1 - p_hat) / 300)
z_star <- qnorm(0.975, mean = 0, sd = 1)
p_hat - z_star * se_p_hat
```

```
## [1] 0.07746206
```

```
p_hat + z_star * se_p_hat
```

```
## [1] 0.1492045
```

Automagic Calculations in R

```
library(mosaic)
confint(binom.test(
  babies_sample$apgar_eq_10,
  conf.level = 0.95,
  ci.method = "wald",
  success = TRUE))
```

```
##   probability of success      lower     upper level
## 1              0.1133333 0.07746209 0.1492046  0.95
```