Math 247: Comparing Proportions from Two Populations (Sections 7.5, 8.4)

Does banding penguins in order to study and track them actually cause harm to the penguins? Claire Saraux, a wildlife researcher, conducted an experiment to answer this question. In 1998, she tagged 200 king penguins at Possession Island in Antarctica. She fitted half the birds with steel bands, and the other half with internal electronic tags. After 10 years of monitoring, Saraux found that just 20 of banded birds had survived the decade, compared to 36 of electronically tagged ones.

Band

No Band (chipped)

What is the research question for this study?

Does banding penguins affect their long-term survival?

What are the two "populations" of interest here?

Banded penguins vs. Unbanded penguins

What are the two samples?

$n_1 = 100$ banded penguins $\quad n_2 = 100$ no band penguins

What is a "success" in this context?

Success = penguin survived for 10 years — note: This is our choice, but it matches the way the data was presented

$x_1 = 20$ survivors (banded) $\quad x_2 = 36$ survivors (unbanded)

We could organize Claire's data in a table:

	Metal Band	No Metal Band
Survived (success/event)	20	36
Didn't survive (failure)	80	64
Total	100	100

What **proportion** of the banded penguins survived? Use proper notation!

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{20}{100} = .20$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{36}{100} = .36$

What **proportion** of the unbanded penguins survived? Use proper notation!

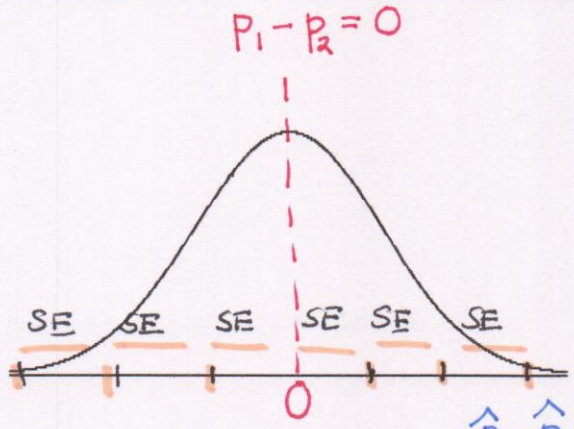
→ Whoops! ☺

What is the difference between the survival **proportions**? Use proper notation!

$\text{difference} = \hat{p}_1 - \hat{p}_2 = .20 - .36 = -.16$

What does the negative sign tell you? It tells us the first number was smaller than the second number. A smaller proportion of banded penguins survived.

Sampling distribution for the difference between two proportions:



$\hat{p}_1 - \hat{p}_2$

Differences in samples will vary.

We can find z-scores now!

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (0)}{SE}$$

Pooled Standard Error:

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Measures Variability Standard Error of Differences

Use this if you assume population proportions are EQUAL (so ZERO difference)

Unpooled Standard Error:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

ignore for now

Now that we have a sampling distribution, we can make an inference about the difference between the population proportions based on the difference between our sample proportions. In laymen's terms, we can see if the data provides evidence that there would be a significant difference in survival rates between banded and unbanded penguins for ALL penguins (not just the penguins in the samples!).

Hypothesis Test for Two Proportions

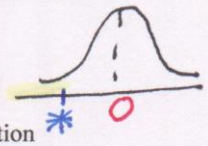
H_0 : The proportion of successes in each of the two populations is the same.

$p_1 = p_2 \Rightarrow \quad p_1 - p_2 = 0$ (the hypothesized difference is zero)

H_a : The proportion of successes of the first **population** is less than the proportion of the second **population**

$p_1 < p_2 \quad \Rightarrow \quad p_1 - p_2 < 0$

Left-tailed Test:

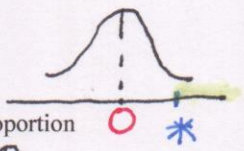


* = $\hat{p}_1 - \hat{p}_2$ = OBSERVED difference between the samples

or the first proportion is more than the second proportion

$p_1 > p_2 \quad \Rightarrow \quad p_1 - p_2 > 0$

Right-tailed Test



or the first proportion different from the second proportion

$p_1 \neq p_2 \Rightarrow \quad p_1 - p_2 \neq 0$

Two-tailed Test

Observed value could be on either side!

We'll use StatCrunch to do the calculations but for reference, here are the formulas:

Use StatCrunch in the HW - will do one by hand, ONCE, in the notes! :)

Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{SE}$ where $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ is the "pooled proportion" and

$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ is the "Pooled Standard Error"

Can we infer from this result that IF half of ALL penguins were banded and the other half of ALL penguins were not banded, that there would be EXACTLY a 16% difference in survival rates? Explain.

No! This data came from samples and we know that samples will vary. The question is, would there be a 16% difference in survival due to just chance, or is that TOO big of a difference?

What should the **difference** be between survival *proportions* if the bands don't affect the penguins' survival?

Difference between the *populations*: ZERO!

$P_{band} - P_{no\ band} = 0$

P = proportion of penguins in the POPULATION who would survive 10 years

Difference between the *samples*:

$\hat{P}_{band} - \hat{P}_{no\ band} = ?$ → This will VARY, but if the bands don't really make a difference, then there shouldn't be a huge difference between the actual sample penguins.

This will jump around but should be close to 0.

Just as we found for a single proportion, if certain conditions are met, then the **differences** between the SAMPLE proportions will be approximately NORMAL in their distribution.

Conditions:

1. **Random and Independent.** Each sample is randomly drawn from its population, and observations are independent from one another.

2. **Large Samples.** Both sample sizes are large enough that there are at least 10 expected successes and at least 10 expected failures in each sample. For this step, we use the "Pooled Proportion" $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ which gives the overall proportion of successes in both samples combined.

3. **Large Populations.** The population size must be at least 10 times sample size, for each sample.

4. (New condition!) **Independent Samples.** The samples themselves are independent from each other.

   For example, if the banded birds and the unbanded birds were deliberately chosen as siblings, then the two samples would not be independent from each other!

not saying whether there's an increase or a decrease, just that the rates are different ≠

Test whether the penguin data suggests there is a statistically significant **difference** in penguin survival rates based on banding. For the hypothesis test, do Step 1 by hand, do Step 2, do Step 3 by hand then using StatCrunch, and do Step 4 by writing a complete conclusion that answers the research question.

Step 1: Hypothesize

$H_0: p_{band} - p_{no\,band} = 0$ There would be ZERO difference in survival rates (proportions) IF we were to band or not band ALL penguins

$H_a: p_{band} - p_{no\,band} \neq 0$ The bands do affect survival! (Maybe help, maybe hurt... were not saying either way)

Step 2: Prepare

Choose significance level: $\alpha = .05$

Choose test:
Two Sample z-Test for Two Proportions

Check conditions:

1. Random samples – assume (unknown)
   Independence WITHIN samples? assume.

2. LARGE samples? Expected successes and failures ≥ 10.

Banded:
Successes: $E = n_1\hat{p} = 100(.28) = 28 \geq 10$ ✓
failures $E = n_1(1-\hat{p}) = 100(.72) = 72 \geq 10$ ✓

Unbanded:
Succ: $E = n_2\hat{p} = 100(.28) = 28 \geq 10$ ✓
fail: $E = n_2(1-\hat{p}) = 100(.72) = 72 \geq 10$ ✓

3. LARGE populations? Unknown, assume there are LOTS of these penguins being studied

4. Independence BETWEEN groups?
   Unknown – assume
   (more about this coming in Chapter 9!)

Parking Lot

Banded:
Successes $x_1 = 20$
observ. $n_1 = 100$
Sample proportion: $\hat{p}_1 = \frac{20}{100} = .20$

No band
Successes $x_2 = 36$
observ. $n_2 = 100$
Sample proportion: $\hat{p}_2 = \frac{36}{100} = .36$

POOLED Success rate

$$\hat{p} = \frac{\hat{p}_1 + \hat{p}_2}{n_1 + n_2} = \frac{56}{200} = .28$$

(28% of all sampled penguins survived)

Step 3: Compute: Find P-value!

$$\text{P-value} = P(\hat{p}_1 - \hat{p}_2 \neq 0)$$

$$= 2 \cdot \text{area in the tail}$$

$$\text{Area in tail} = 2 \cdot P(z \leq -2.52)$$

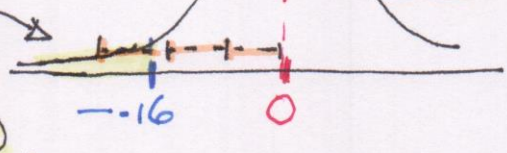
$$= 2 \cdot (.00587)$$

$$= .01174$$

$$z = \frac{\text{observed} - \text{center}}{SE} = \frac{-.16 - 0}{.0634} = -2.52$$

NULL

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = .0634$$



Using StatCrunch: More complete instructions are at the end of Chapter 8

Stat → →Proportion Stats →Two Sample →With Summary
Fill in the number of successes and the number of observations for each sample
Select the "P-value plot" under "Optional Graphs and Tables"

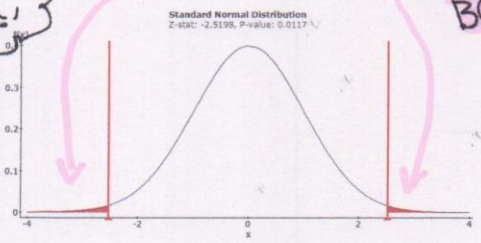
observed SE z

Hypothesis test results:

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	20	100	36	100	-0.16	0.0634	-2.51976	0.0117

banded not banded

In the homework, you'll do the problems using StatCrunch, but I want you to know what each value in the table represents!



area in BOTH tails, combined because $H_a \neq$

Step 4: Interpret

P-value = .0117 < .05 = α

Reject H_0, accept H_a.

We have evidence that banding made a significant difference in survival rates AT THE .05 level of significance. (If α were .01 we would Not reject H_0!)

The banded penguins were significantly less likely to ~~survive~~ survive for 10 years, as compared to the unbanded penguins.