Proportions

Introduction to Statistics
Carl von Ossietzky Universität Oldenburg
Fakultät III - Sprach- und Kulturwissenschaften

Binomial setting

- **Example**: 2500 adults in the US were asked whether they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming". 1650 respondents agreed. (Example from Moore & McCabe, 2006).
- **Sample proportion**:

$$\hat{p} = \frac{X}{n} = \frac{1650}{2500} = 0.66$$

- There are a fixed number of n observations. The n observations are all independent.
- Each observation falls into one of just two categories, which for convenience we call "success" and "failure."
- The probability of a succes, call it p, is the same for each observation.

Binomial distribution

- **Example**: toss a coin n times. Each toss gives either a head or a tail.
- Outcomes of successive tosses are independent!
- If we call heads a succes, then p is the probability of a head and remains the same as long as we toss the same coin.
- X is the number of heads we count. Its distribution is determined by n and p.
- The distribution of the count X of successes in the binomial setting is called the **binomial distribution** with parameters n and p.
- The parameter n is the number of observations, and p is the probability of a success on any one observation.
- The possible values of X are the whole numbers from 0 to n.
- As an abbreviation, we say that X is $B(n, p)$.

Sampling distribution of a count

- A population contains proportion p of successes.
- If the population is much larger than the sample, the count X of successes in an SRS of size n has approximately the binomial distribution $B(n, p)$.
- The accuracy of this approximation improves as the size of the population increases relative to the size of the sample.
- As a rule of thumb, we will use the binomial sampling distribution for counts when the population is at least 20 times as large as the sample.

Sampling distribution of a count

- **Example**: tossing a coin. The chance to get a head is 0.5. You toss the coin 10 times. What is the chance to have only three heads?
- Go to `http://www.vassarstats.net/`, choose Distributions, Binomial Distributions. Click on Continue and enter n=10, p=0.50.
- Look for row k=3 and column $Exact\ Probability$. We find $P(X = 3) = 0.1172$ or 11.72%.
- This probability can be found in three steps. First find the number of ways of arranging k successes among n observations. Calculate the **binomial coefficient**:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where the **factorial** $n!$ is:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$$

Sampling distribution of a count

- So:

$$\binom{n}{k} = \frac{10!}{3!(10-3)!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$$

- Second we calculate the probability to get a sequence with k successes (heads) and $n - k$ failures (tails) in a particular order:

$$p^k(1-p)^{n-k}$$

- So the probability to get 3 heads and 7 tails in a particular order is:

$$0.5^3(1-0.5)^{10-3} = 0.0010$$

Sampling distribution of a count

- Third we calculate the **binomial probability**:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- In our case this becomes:

$$P(X = 3) = (120)(0.0010) = 0.1172$$

- Other question: What is the chance to get at most three heads? Or: what is the chance that at most 30% of the tosses have a head?

$$\begin{array}{lcccccccccc} P(X \leq 3) & = & P(X=0) & + & P(X=1) & + & P(X=2) & + & P(X=3) & = & \\ & & 0.0010 & + & 0.0098 & + & 0.0439 & + & 0.1172 & = & 0.1719 \end{array}$$

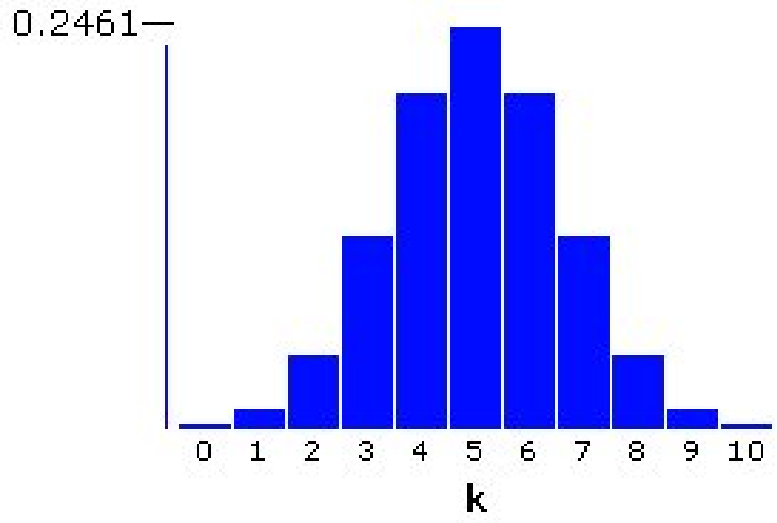
or 17.19%.

Sampling distribution of a count

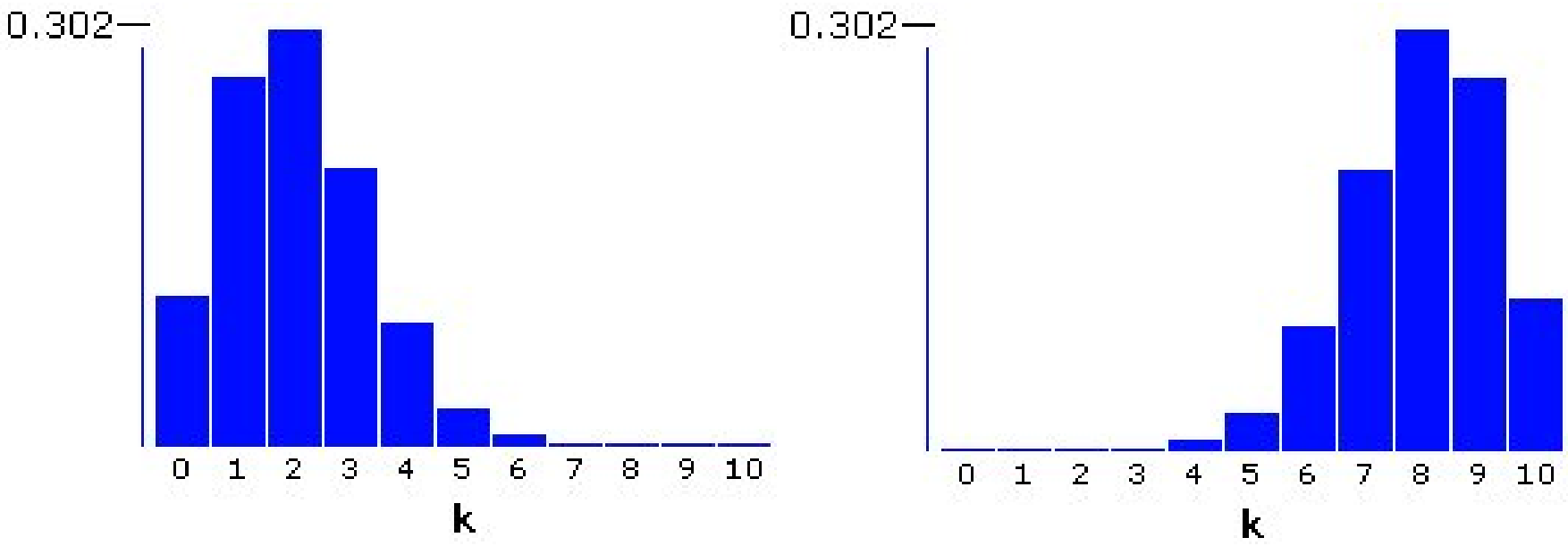
- Find the probabilities with **VassarStats: Website for Statistical Computation**:

  `http://faculty.vassar.edu/lowry/VassarStats.html`

- Select Distributions, select Binomial Distributions, enter n=10 and p=0.5. You get a probability histogram and a table.



Sampling distribution of a count



Two histograms with $n = 10$. Left: $p = 0.2$, right: $p = 0.8$.

Sampling distribution of a count

[Results are rounded to a maximum of 12 decimal places.]

		Cumulative Probability	
k	**Exact Probability**	From Left to Right Sum of Exact Probabilities for 0 through k, inclusive	From Right to Left Sum of Exact Probabilities for k through 10, inclusive
0	0.0009765625	0.0009765625	1.0
1	0.009765625	0.0107421875	0.9990234375
2	0.0439453125	0.0546875	0.9892578125
3	0.1171875	0.171875	0.9453125
4	0.205078125	0.376953125	0.828125
5	0.24609375	0.623046875	0.623046875
6	0.205078125	0.828125	0.376953125
7	0.1171875	0.9453125	0.171875
8	0.0439453125	0.9892578125	0.0546875
9	0.009765625	0.9990234375	0.0107421875
10	0.0009765625	1.0	0.0009765625

Mean and standard deviation of a count

- A binomial variable X is the count of successes in n independent observations that each have the same probability p of succes.
- If a count X has the binomial distribution $B(n, p)$, then:

$$\mu_X = np$$

and:

$$\sigma_X = \sqrt{np(1-p)}$$

- In our example:

$$\mu_X = (10)(0.5) = 5$$

and:

$$\sigma_X = \sqrt{(10)(0.5)(1-0.5)} = 1.5811$$

Mean and standard deviation of a sample proportion

- Let $\hat{p}$ be the sample proportion of successes in an SRS of size n drawn from a large population having population proportion p of successes.
- The the mean of $\hat{p}$ is:

$$\mu_{\hat{p}} = p$$

and the standard deviation of $\hat{p}$ is:

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

- In our example:

$$\mu_{\hat{p}} = 0.5$$

and:

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.5)(1-0.5)}{10}} = 0.1581$$

Normal approximation for counts and proportions

- Draw an SRS of size n from a large population having population proportion p of successes.
- Let X be the **count** of successes in the sample and $\hat{p} = X/n$ the sample **proportions** of successes.
- When n is large, the sampling distributions of these statistics are approximately normal. X is approximately:

$$N\left(np, \sqrt{np(1-p)}\right)$$

and $\hat{p}$ is approximately:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- As a rule of thumb, we will use this approximation for values n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Large-sample confidence interval for a population proportion

- The standard deviation $\sigma_{\hat{p}}$ depends upon the the parameter p. When constructing a confidence interval, p is unknown.
- To estimate this standard deviation using the data, we replace p in the formula by the sample proportion $\hat{p}$.
- **Standard error**: standard deviation estimated from sample data. The standard error of $\hat{p}$ is:

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

Large-sample confidence interval for a population proportion

- To construct **an approximate level C confidence interval** for p we first catch the central area C under a normal curve.
- We must find the number z^* such that any normal distribution has probability C within $\pm\ z^*$ standard deviations of its mean.
- There is probability C that p lies between:

$$\hat{p} - z^* SE_{\hat{p}} \quad and \quad \hat{p} + z^* SE_{\hat{p}}$$

- Use this interval for 90%, 95% or 99% confidence when the number of successes and the number of failures are both at least 15.

Large-sample confidence interval for a population proportion

- **Example**: 2500 adults in the US were asked whether they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming". 1650 respondents agreed.
- The sample proportion is:

$$\hat{p} = 1650/2500 = 0.66$$

- The standard error of $\hat{p}$ is:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.66)(0.34)}{2500}} = 0.009474$$

Large-sample confidence interval for a population proportion

- We calculate the 95% confidence interval. Go to `http://www.vassarstats.net/`, choose Distributions, choose Normal Distributions. Enter mean=0 and standard deviation=1 (standard normal distribution).
- If C=0.95, than p=1-C=0.95=0.05. Find p=0.05 in the table at the right and look in the column $2\ tail$. We find 1.96.
- Therefore z^*=1.960. Then the unknown population proportion p lies between:

$$\hat{p} - z^* SE_{\hat{p}} \quad and \quad \hat{p} - z^* SE_{\hat{p}}$$

is:

$$0.66 - (1.960)(0.009474) \quad and \quad 0.66 + (1.960)(0.009474)$$

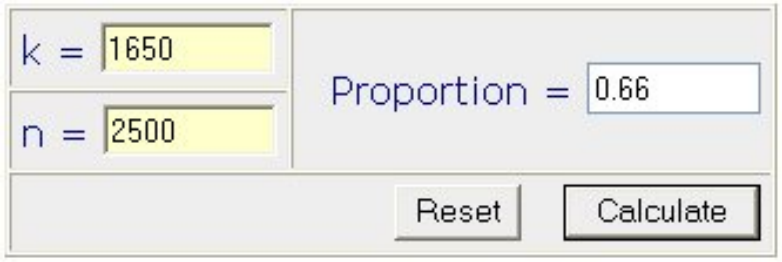
is:

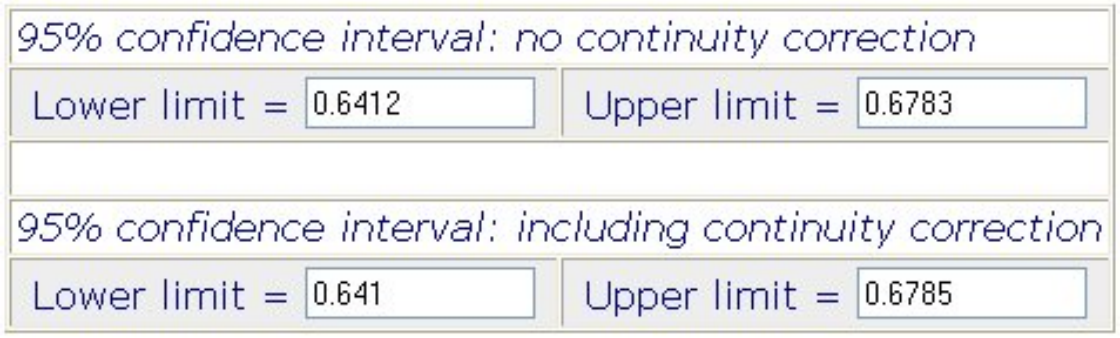
$$0.64 \quad and \quad 0.68$$

- We estimate with 95% confidence that between 64% and 68% of the US adults will agree with "I like buying new clothes, but shopping is often frustrating and time-consuming".

Large-sample confidence interval for a population proportion

- In **VassarStats**: select Proportions, select The Confidence Interval of a Proportion, and enter k and n only:





Large-sample confidence interval for a population proportion

- The level C confidence interval for a proportion p will have margin of error approximately equal to a specified value m when the sample satifies:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

- Here z^* is the critical value for confidence C, and p^* is a **guessed** value for the proportion of successes in the future sample.
- Two ways to get p^*:
  - Use sample estimate from pilot study or similar studies done earlier.
  - Use $p^* = 0.5$. Because margin of error is largest in that case, this choice gives a sample size that is somewhat larger than we really need for the confidence level we choose.

Large-sample confidence interval for a population proportion

- We return to our shopping example. How large should the sample be to obtain a 95% confidence interval and a margin of error less than or equal to 3% or 0.03?
- Earlier we found that for a 95% confidence, z^*=1.960. We choose $p^* = 0.5$. Required sample size:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) = \left(\frac{1.960}{0.03}\right)^2 (0.5)(0.5) = 1067.1111$$

- Always round up, rounding down would give a margin of error slightly greater than 0.03: n=1068.

Large-sample significance test for a population proportion

- In the sample of 2500 US residents we found that 1650 adults agreed with "I like buying new clothes, but shopping is often frustrating and time-consuming".
- Assume that in a larger national survey it was found that 68% agreed. Is our sample compatible with this larger survey?
- We test the null hypothesis:

$$H_0\text{: } p = 0.68$$

and choose the two-sided alternative hypothesis:

$$H_a\text{: } p \neq 0.68\text{, the } p\text{-value is } P(Z \neq z)$$

Large-sample significance test for a population proportion

- Right-sided alternative:

$$H_a\text{: } p > 0.68\text{, the } p\text{-value is } P(Z \geq z)$$

- Left-sided alternative:

$$H_a\text{: } p < 0.68\text{, the } p\text{-value is } P(Z \leq z)$$

- We know that:

$$\mu_{\hat{p}} = \hat{p} = 0.66$$

and

$$\sigma_{\hat{p}} = \sqrt{\frac{(p)(1-p)}{n}} = \sqrt{\frac{(0.68)(0.32)}{2500}} = 0.0093$$

Large-sample significance test for a population proportion

- The test statistic is:

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.66 - 0.68}{0.0093} = -2.14$$

- $P(Z \neq -2.14) = P(Z \leq -2.14) + P(Z \geq 2.14) = 2P(Z \leq -2.14)$.
- Go to `http://www.vassarstats.net/tabs.html` and choose *z to P*. Enter $z = -2.14$ and click on Calculate. Look in the row *two-tailed for* $\pm z$. We find: p=0.0324.
- We choose $\alpha = 0.05$ which is higher than the p-value of 0.0324. We reject H_0 and accept H_a: $p \neq 0.68$.
- Conclusion: our sample is not compatible with the results of the larger survey.

A paired-sample problem

- Tanja Gaustad (2004) implemented two methods to determine intended word senses in text. She tested her software on 55,000 examples of words for which the disambiguation was known.
- Note that the two methods were applied to the same data (paired data).
- When she contrasted the methods, there were three possible outcomes:
  1. The methods agreed (these cases were discounted).
  2. The first method was right, and the second wrong.
  3. The second method was right, and the first wrong.
- In a typical contrast the methods agreed in all but 500 cases.
- We reason from a background assumption that the methods differ on the basis of chance: H_0: $p = 0.5$
- How many more examples does the better method need in order to be significantly better?

A paired-sample problem

- We calculate:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{500}} = 0.0224$$

- If $\alpha = 0.05$, then z=1.645. The test statistic is:

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - 0.5}{0.0224} = 1.645$$

- So:

$$\hat{p} = 1.645 \times 0.0224 + 0.5 = 0.5368$$

- Increase : 0.5368-0.50=0.0368 or 3.7%; 3.7% of 500 is 19 (rounded up).
- A method that only got 19 more examples right than a competitor is thus better to a statistically significant degree.

Large-sample confidence interval for comparing two proportions

- Because comparitive studies are so common, we often want to compare the proportions of two groups that have some characteristic.
- Choose an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes.
- The estimate of the difference in the population proportions is:

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

- When both sample sizes are sufficiently large, the sampling distribution of the difference D is approximately normal.
- The **standard error of D** is:

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

Large-sample confidence interval for comparing two proportions

- To construct **an approximate level C confidence interval** for p we first catch the central area C under a normal curve.
- We must find the number z^* such that any normal distribution has probability C within $\pm\ z^*$ standard deviations of its mean.
- There is probability C that $p_1 - p_2$ lies between:

$$D - z^* SE_D \quad and \quad D + z^* SE_D$$

- Use this interval for 90%, 95% or 99% confidence when the number of successes and the number of failures in both samples are all at least 10.

Large-sample confidence interval for comparing two proportions

- **Example**: Nynke van den Bergh studies children acquiring Frisian. There are two groups:
  - children who hear only Frisian at home and in child-care settings.
  - children who hear Frisian at home and Dutch in child-care settings.
- The question is whether the mixed setting will lead to more interference errors—errors whether the child uses a Dutch pattern instead of a Frisian one.
- Van den Bergh's data for kids 5 years, 11 months old:

Setting	**Correct**	**Incorrect**
Pure Frisian	85 (97.7%)	2 (2.3%)
Mixed	167 (89.8%)	19 (10.2%)

Large-sample confidence interval for comparing two proportions

- Let's find a 95% confidence interval for the difference between the proportion of inference errors of children in a purley Frisian environment and the proportion of inference errors of children in a mixed environment:

$$D = \hat{p_1} - \hat{p_2} = 0.023 - 0.102 = -0.079$$

- The standard error of D is:

$$SE_D = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} = \sqrt{\frac{(0.023)(1-0.023)}{85+2} + \frac{(0.102)(1-0.102)}{167+19}} = 0.0274$$

Large-sample confidence interval for comparing two proportions

- We calculate the 95% confidence interval. For 95% confidence, we found that z^*=1.960. Then $p_1 - p_2$ lies between:

$$D - z^* SE_D \quad and \quad D + z^* SE_D$$

is:

$$-0.079 - (1.96)(0.0274) \quad and \quad -0.079 + (1.96)(0.0274)$$

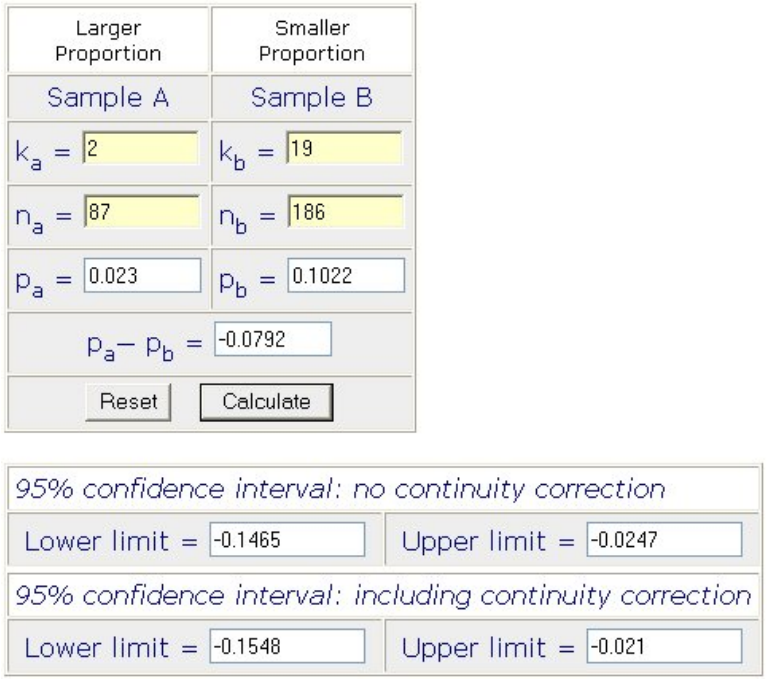
is:

$$-0.1327 \quad and \quad -0.0253$$

- With 95% confidence we can say that the difference in proportions is between -0.1327 and -0.0253. Alternatively we can report that 'pure Frisian children' are 7.9% **less** likely to make inference errors than 'mixed Frisian children' with a 95% margin of error of 5.4%.

Large-sample confidence interval for comparing two proportions

- In **VassarStats**: select Proportions, select The Confidence Interval for the Difference Between Two Independent Proportions, and enter k_1, n_1, k_2 and n_2 only:



Significance test for comparing two proportions

- Van den Bergh's null hypothesis is of course, that there is **no** difference in the proportion of incorrect inflections in the two populations (of expressions of inflection among children from the purely Frisian environment on the one hand as opposed to the children from the mixed environment on the other). Her alternative hypothesis is that the children from the mixed environment show more errors due to interference.
- We test the null hypothesis:

$$H_0\text{: } p_F = p_M$$

and choose the left-sided alternative hypothesis:

$$H_a\text{: } p_F < p_M\text{, the } p\text{-value is } P(Z \leq z)$$

where p_F is the error percentage of children in the purely Frisian environment, p_M the error percentage of children in a mixed environment.

Significance test for comparing two proportions

- Right-sided alternative:

$$H_a\text{: } p_F > p_M\text{, the } p\text{-value is } P(Z \geq z)$$

- Two-sided alternative:

$$H_a\text{: } p_F \neq p_M\text{, the } p\text{-value is } P(Z \neq z)$$

- We wish to assume that there is **no** difference in the proportions in the two populations (the population of kids' expressions in the pure setting and their population in the mixed setting), and ask how likely these samples are given that assumption.

Significance test for comparing two proportions

- To test the null hypothesis we compute the z **statistic**:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{D_p}}$$

- The **pooled standard error** is:

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

- The pooled standard error uses the **pooled estimate** $\hat{p}$, which combines, or pools, the information from both samples:

$$\hat{p} = \frac{number\ of\ successes\ in\ both\ samples}{number\ of\ observations\ in\ both\ samples} = \frac{X_1 + X_2}{n_1 + n_2}$$

Significance test for comparing two proportions

- In our example:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{2 + 19}{85 + 2 + 167 + 19} = 0.0769$$

- The **pooled standard error** is:

$$SE_{Dp} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.0769)(1-0.0769)\left(\frac{1}{85+2} + \frac{1}{167+19}\right)} = 0.0346$$

- The z statistic is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{Dp}} = \frac{0.023 - 0.102}{0.0346} = -2.28$$

Significance test for comparing two proportions

- Go to `http://www.vassarstats.net/tabs.html` and choose z *to* P. Enter $z = -2.28$. Look in the row *one-tailed for -z*. We find 0.0113.
- We choose $\alpha = 0.05$ which is higher than the p-value of 0.0113. We reject H_0 and accept H_a: $p_F < p_M$.
- Conclusion: The error percentage of children in the purely Frisian environment is significantly lower than the error percentage of children in a mixed environment

Significance test for comparing two proportions

- In **VassarStats**: select Proportions, select Significance of the Difference Between Two Independent Proportions and enter k_1, n_1, k_2 and n_2 only: