

## Topic#5 Estimation of a Proportion

Three examples:

- \* A congressional leader investigating the merits of an 18 year old voting age may want to estimate the proportion of the potential voters in the district between the ages of 18 and 21.
- \* A marketing research group may be interested in the porportion of the total sales market in diet preparations that is attributable to a particular product. That is, what percentage of sales is accounted for by a particular product?
- \* A forest manager may be interested in the proportion of trees with a diameter of 12 inches or more.

All these examples exhibit a characteristic of the binomial experiment - that is, an observation either does or does not belong to the category of interest. (two-ness)

We denote the population proportion by  $p$  and the sample proportion by  $\hat{p}$ .

**Estimator of the population proportion  $p$ :**

$$\hat{p} = \frac{\text{Number of elements with the specified characteristic}}{\text{Total number in the sample}}$$

**Estimated variance of  $\hat{p}$ :**

$$\widehat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1} \left( \frac{N-n}{N} \right) \quad \text{where } q = 1 - p$$

**Bound on the error of estimation:**

$$2\sqrt{\widehat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1} \left( \frac{N-n}{N} \right)}$$

Example: A simple random sample of  $n = 100$  college seniors was selected to estimate (1) the fraction of  $N = 300$  seniors going on to grad school and (2) the fraction of students that have held part time jobs during college. Out of our sample of 100, 15 are going on to grad school and 65 held part-time jobs.

A. Estimate the proportion of seniors going on to grad school.

$$\hat{p} = \frac{15}{100} = 0.15$$

$$2\sqrt{\widehat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1} \left( \frac{N-n}{N} \right)} = 2\sqrt{\frac{(.15)(.85)}{99} \left( \frac{300-100}{300} \right)} = 2(0.0293) = 0.059$$

We estimate that 0.15 (or 15%) of the seniors plan to attend graduate school, with a bound of error of estimation equal to 0.059 (or 5.9%).

B. Estimate the proportion of seniors who have had a part time job sometime during their college careers. (YOUR TURN! Bound for error should be 0.078 or 7.8%)

**Sample size required to estimate  $p$  with a bound on the error of estimation B:**

$$n = \frac{Npq}{(N-1)D + pq} \text{ where } q = 1 - p \text{ and } D = \frac{B^2}{4}$$

We are trying to get the sample size needed to estimate p and yet, p is in the equation!  
 We can sometimes get an estimate of p from a similar study. We can do a small survey to get an estimate of p. If there is no estimate for p, we can use p = 0.5 (worst case) but our sample will likely be larger than it needs to be.

Example: Student government leaders at a college want to conduct a survey to determine the porportion of students who favor the proposed honor code. Because interviewing N = 2000 students in a reasonable length of time is almost impossible, determine the sample size (number of students to be interviewed) needed to estimate p with a bound of error of estimation of magnitude B = 0.05.

A. Assume no prior information is available to estimate p.

$$D = \frac{B^2}{4} = \frac{(0.05)^2}{4} = 0.000625$$

$$n = \frac{Npq}{(N-1)D + pq} = \frac{(2000)(0.5)(0.5)}{(1999)(0.000625) + (0.5)(0.5)} = \frac{500}{1.499} = 333.56 = 334$$

B. Assume a similar study at another school showed p to be estimated at 0.25 (or 25%).

$$D = \frac{B^2}{4} = \frac{(0.05)^2}{4} = 0.000625$$

$$n = \frac{Npq}{(N-1)D + pq} = \frac{(2000)(0.25)(0.75)}{(1999)(0.000625) + (0.25)(0.75)} = \frac{375}{1.4369} = 260.978 = 261$$

Example: Student government leaders at a college want to conduct a survey to determine the porportion of students who feel that the student union building currently serves their needs. Because interviewing N = 2000 students in a reasonable length of time is almost impossible, determine the sample size (number of students to be interviewed) needed to estimate p with a bound of error of estimation of magnitude B = 0.07.

A. Assume no prior information is available to estimate p.

B. A similar study done last year shows that approximately 60% of the students felt that the student union served their needs.

## Topic#6 Comparing Estimates

We often want to compare the estimates of two parameters. Two examples:

\* The mean incomes of two ethnic groups over the past year can be compared by looking at the difference between the sample means for random samples of incomes of two groups.

\* Whether the republicans are gaining on the Democrats in a congressional race can be assessed by looking at the difference between the proportions voting Republican for two pools conducted a few weeks apart.

**For the mean or the propotion:**

$$E(y_1 - y_2) = E(y_1) - E(y_2)$$

$$V(y_1 - y_2) = V(y_1) + V(y_2) - 2cov(y_1, y_2)$$

If  $y_1$  and  $y_2$  are independent, then  $cov(y_1, y_2) = 0$ .

Example: Fish absorb mercury through their gills, and too much mercury makes the fish unfit for human consumption. In 1994 the state of Maine issued a health advisory warning that people should be careful about eating fish from Maine lakes because of the high levels of mercury. Before the warning, data on the status of Maine lakes were collected by the U.S. Environmental Protection Agency (EPA) working with the state. Fish were taken from a random sample of lakes and their mercury content was measured in parts per million (ppm). The table#1 attached shows the actual data from a random sample of 35 lakes.

Type 1: Oligotrophic - balanced between decaying vegetation and living organisms.

Type 2: Eutropic - high decay rate and little oxygen.

Type 3: Mesotrophic - between the other two states.

The table also shows whether the lake is formed behind a dam.

We would like to:

- A. Compare types 1 and 2 and estimate the difference in mean mercury levels.
- B. Decide if the mean mercury level for type 2 differs from type 3.
- C. Compare types 1 and 3 and estimate the difference in mean mercury levels.

The following has been computed using the data in Table#1:

Type	Count	Mean	Median	Standard Deviation
1	4	0.22	0.20	0.103
2	15	0.74	0.68	0.583
3	16	0.50	0.44	0.272

**A. Compare types 1 and 2 and estimate the difference in mean mercury levels.**

$$\begin{aligned}
 (y_1 - y_2) \pm 2\sqrt{[V(y_1) + V(y_2)]} &= (0.22 - 0.74) \pm 2\sqrt{\left[\frac{0.103^2}{4} + \frac{0.583^2}{15}\right]} \\
 &= -0.52 \pm 0.32
 \end{aligned}$$

The true difference could be anywhere between -0.84 and -0.20.

The total number of lakes is quite large, so the finite population corrections are ignored.

**B. Decide if the mean mercury level for type 2 differs from type 3.**

$$\begin{aligned}
 (y_2 - y_3) \pm 2\sqrt{[V(y_2) + V(y_3)]} &= (0.74 - 0.50) \pm 2\sqrt{\left[\frac{0.583^2}{15} + \frac{0.272^2}{16}\right]} \\
 &= 0.24 \pm 0.33
 \end{aligned}$$

The resulting interval (-0.09, 0.57) covers zero, which implies that there is no significant evidence of a difference of mean mercury content for these two types of lakes!

**C. Compare types 1 and 3 and estimate the difference in mean mercury levels.  
Your turn!**

**When comparing means**, we consider only the independent sample case because the dependent case becomes too complicated to handle at this level.

**When comparing proportions**, however, a commonly occurring dependent situation does have a rather simple solution.

We will use the following formulas for proportion:

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{n} + \frac{p_2q_2}{n} + 2\frac{p_1p_2}{n}$$

Example: A Time/Yankelovich poll of 800 adult americans carried out in 1994 asked "Should smoking be banned from workplaces, should there be special smoking areas, or should there be no restrictions?" The results are summarized below for 600 nonsmokers and 200 smokers:

	Nonsmokers (%)	Smokers (%)
Banned	44	8
Special Areas	52	80
No Restrictions	3	11

**A. Find the true difference between the proportions choosing "banned". (independent)**

$$(0.44 - 0.08) \pm 2\sqrt{\left[\frac{(0.44)(0.56)}{600} + \frac{(0.08)(0.92)}{200}\right]} = 0.36 \pm 0.06$$

The true difference is between 30% and 42%.

**B. Find the true difference between the proportions nonsmokers choosing "banned" and "special areas". (dependent - if one is small, the other must be large!)**

$$(0.52 - 0.44) \pm 2\sqrt{\left[\frac{(0.52)(0.48)}{600} + \frac{(0.44)(0.56)}{600} + 2\frac{(0.44)(0.52)}{600}\right]} = 0.08 \pm 0.08$$

The true difference is between 0% and 16%.

There is no evidence that these populations differ.

**C. Find the true difference between the proportions choosing "no restrictions". (Your turn!)**

**D. Find the true difference between the proportions smokers choosing "banned" and "special areas". (Your turn!)**

**Table#1: Mercury content in Maine lakes according to type and dam:**

Mercury (ppm)	Lake Type	Dam, 1=yes, 0=no
1.050	2	1
0.230	2	1
0.100	3	0
0.770	2	1
0.910	2	1
0.250	2	1
0.130	1	1
0.290	2	0
0.410	3	1
0.210	3	0
0.940	2	0
0.360	1	1
1.220	2	0
0.240	1	1
0.900	3	0
2.500	2	1
0.340	3	0
0.400	3	1
0.450	2	1
1.120	3	1
0.320	2	0
0.370	3	0
0.540	3	0
0.860	3	0
0.770	2	0
0.670	3	0
0.600	3	1
0.680	2	1
0.220	3	1
0.470	3	1
0.370	3	1
0.290	2	0
0.430	2	1
0.160	1	0
0.490	3	0

Source: R. Peck, L. Haugh, and A. Goodman, 1998, Statistics Case Studies, ASA-SIAM, 1-14.

## Topic#7 Summary

**The objective** of the sample survey is to make inferences about one or more of the population parameters from the information contained in a sample. Two factors affect the quantity of information in a given investigation. The first is the sample size. The larger the sample size, the more information we expect to obtain about the population. The second factor is the amount of variation in the data. Variation can be controlled by the design of the sample survey.

**Simple Random Sampling** is the simplest type of survey design. This design does not attempt to reduce the effect of data variation on the error of the estimation. A simple random sample size of  $n$  occurs if each sample of  $n$  elements from the population has the same chance of being selected. Random number tables are quite useful in determining the elements to be included in a simple random sample.

**In estimating a population mean  $\mu$  and total  $T$** , we use the sample mean and the sample total respectively. Both estimators are unbiased, that is, the expected value of the sample mean is approximately equal to the population mean. (Same holds true for the total.) The estimated variance and bound for the error of estimation were computed for both estimators.

Sometime during the actual design of the survey, the experimenter must decide how much information is desired, that is, how large a bound on the error of estimation can be tolerated. **Sample size requirements** have been calculated for the **mean and the total**.

The third parameter estimated was the **population proportion  $p$** . The properties of estimating proportion, bound for error of estimation, and calculation of sample size needed were related to the properties of the mean.

Sometimes it is important to **make comparisons among means and proportions** by estimating differences. The variances of these differences are estimated for independent means. The variances of these differences are estimated for independent and dependent proportions.