

Taking the Sting Out of Statistics, Part 2: Statistical Inference

Elana Broch (ebroch@princeton.edu)

Stokes Library for Public and International Affairs
& the Coale Population Research Collection
Princeton University

Introduction/motivation

Standard Normal Distribution and Area under the Curve

Sampling Variability

Random Sampling Distributions

Sampling distribution of proportion

- **Proportion**=number of people with the attribute divided by the total number of people (e.g., proportion of males).
 - A proportion is a number between zero and 1. 70% is equivalent to .70 or .7
- Exercise on sampling distributions

Confidence Interval for proportion

Margin of error

Hypothesis testing for a proportion (statistical significance)

Conclusions: extending these concepts to other situations (comparing two groups' means, chi-square, ANOVA, correlation).

A Fantastic Reference (referenced in Handout as M&M)

Moore, D.S. and McCabe, G.P. (2003) *Introduction to the practice of statistics*, 4th ed. New York: W.H. Freeman and Co.

Earlier we talked about descriptive statistics. We learned a lot of terminology, and discussed the importance of looking at the graph of data in addition to statistics such as the mean or the median. We assumed we were summarizing the group of interest. We looked at mean and median salary data from SLA, and graphs of educational attainment and percent of Hispanic adults in a state.

Now we'll take that a step further and explore the really powerful part of statistics--inferential statistics. In many real world application we want to be able to generalize to a population, based on a sample from that population.

Here are some situations that you're probably familiar with that use inferential statistics.

- Predicting election results

- Measuring the President's "approval rating"

- Testing the efficacy of a new drug or treatment protocol

- Unemployment rate

- Other examples?

In these examples, we are interested in a very large group of people but base our findings on a **sample** of that larger group (i.e., the **population** of interest).

By contrast, there are a few situations where inference is not the cornerstone of the approach because you have (in theory) measured every person you're interested in. Census (short form, not long form, questions); mean GRE scores; birth and death records.

It is usually impractical to measure the population of interest?

- Why? Time, money. Or, perhaps the treatment is risky.

We do this by selecting a sample and computing a statistic based on that sample (e.g., proportion of males or mean birthweight.) We then use the sample statistics to generate a **sampling distribution**. Here we are thinking about only one variable at a time.

Review of distributions

A **variable** varies. Examples: gender, percent Hispanics living in a state, educational attainment of adults 25 and older, reaction to hearing the word “statistics”. Variable is one column in data file.

The **distribution** of a variable tells us what values it takes and how often it takes these values [M&M, p. 5]. Can be a table or a graphy.

A distribution can be displayed on a graph with values of the variable along the x- (horizontal) axis. The y-axis has the **frequency** (number) or **relative frequency** (proportion or percent).

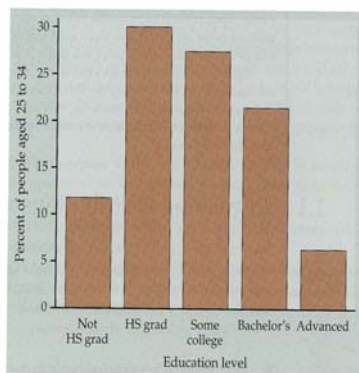


FIGURE 1.1(a) Bar graph of the educational attainment of people aged 25 to 34 years.

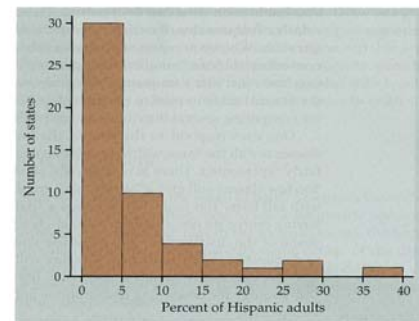


FIGURE 1.4 Histogram of the percent of each state's adult residents who identified themselves as Hispanic in the 2000 census.

Percent of people with advanced degrees is what? As a proportion equals what? Total of percents = 100 or 1.0 as a proportion.

Game plan for Statistical Inference

There is a population with a value on a variable of interest, p

We get a value of that variable for a random sample, \hat{p}

We generate a Sampling Distribution of that variable given the sample size we used and the value of \hat{p} .

So we can make a statistical inference

- Establish a Confidence Interval, or
- Perform hypothesis testing (statistical significance)

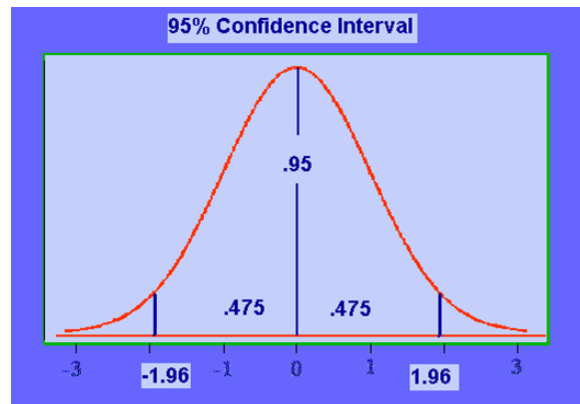
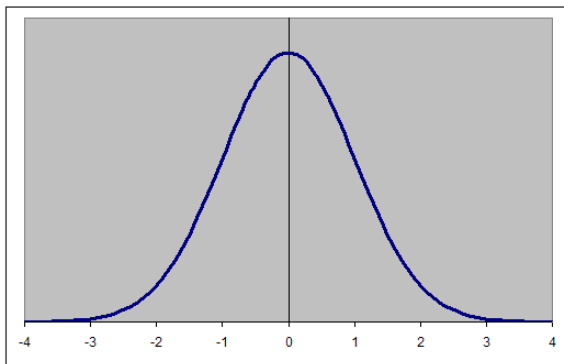
Repeat for additional variables, if necessary.

Standard Normal Distribution

A very important distribution is the Standard Normal Distribution (a.k.a., bell curve). Unlike the educational attainment or Hispanic adults by states which were based on data, this distribution is theoretical (nothing really looks exactly like this).

Like the graph of educational attainment, total area=100% or 1.0, as a proportion.

In fact, mathematicans have worked out the exact area for the different regions of the normal curve, based on the assumption that the entire area equals 1.0.



Accessed online from <http://academics.rmu.edu/faculty/clavijo/standa11.gif>

Half the area is above zero, half is below zero. Much of the area is between -1 and 1. Very little area is below -2 or above 2. These areas can be seen in the figure on the right.

.68 of the area is between -1 and 1

.95 of the area is between -1.96 and 1.96

Sampling Variability

Imagine we were interested in estimating the population of males and females born in the United States in 2004. We actually know this number from birth records (ignoring births not reported to NCHS), but we'll assume we don't. We would take a random sample of births, and record whether the person is male or female, and figure out the proportion of males in this sample. This number will range from 0 to 1. Is it likely to be zero, one, .2, .5?

If the sample has 60% (or .6) males, how likely is it that the underlying population was 50% male? $\hat{p} = .6$ The answer depends on the size of the sample you're working with.

Key concept: sample proportion may underestimate, overestimate, or exactly equal the population proportion. Although we know the sample proportion, we don't know if it's underestimating, overestimating, or equaling the population proportion. If the sample is large enough, it should be a reasonable estimate of the population proportion.

We want a mechanism that takes into account the sample size and the standard deviation of the population to get a sense of how much the sample might underestimate, overestimate the population value. One way to do this would be to keep taking samples of the same size and kept track of the sample proportion of each of them we would get what's called an empirical Random Sampling Distribution (see powerpoint slides). Fortunately, statistical theory saves us from that repeated sampling, by giving us a theoretical (i.e., not empirically derived) Random Sampling Distribution.

To give you a sense of what a sampling distribution is, I have an exercise.

Baby Shoes

12 volunteers ("cringe when you hear stats")

Post-its with sample proportions

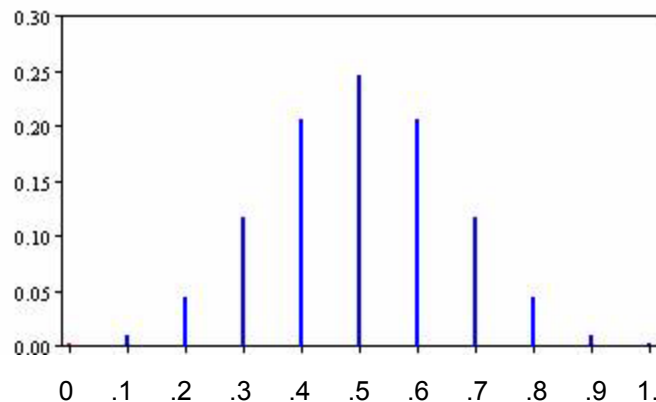
See powerpoint babies and Julia's slides

Random Sampling Distributions

A random sampling distribution is a theoretical, relative frequency distribution of all possible sample values (e.g., the proportion of males in a sample).

This graph represents the (theoretical) relative frequency distribution if you were to take a large number of samples of size 10 from a population where the underlying proportion of males was .5

Relative frequency distribution of samples of size 10 from a population with a proportion of .5
Created using applet at http://psych.rice.edu/online_stat/analysis_lab/binomial_dist.html



The proportion of males in the sample is .5 in 25% (.25 on the y-axis) of the samples.

The proportion of males in the sample is .4 in 20% (.20 on the y-axis) of the samples.

The proportion of males in the sample is .6 in 20% (.20 on the y-axis) of the samples.

Almost none of the samples would have 0 or 10 males (0 or 1.0)

If you had a population with a proportion of males=.5, what proportion of samples would have proportion of males=.6 ? It is certainly possible that a sample proportion of .6 came from a population in which the proportion of males was .5, as seen in the graph above.

Summarize RSDM

Using sampling distributions to establish a confidence interval

You are all familiar with confidence intervals. You've heard it said that the results of a poll have a margin of error of 3 percentage points. A confidence interval is a range of possible values for the population proportion based on the sample proportion you observed.

Pollsters try to get an adequate sample size so that their results are within a certain margin of error. But given the cost of polling or interviewing, keeping the number as small as possible. This is the same as is done with the CPS (Current Population Survey).

For 95% confidence interval: sample proportion \pm 1.96 times the standard deviation of sampling distribution, where \pm means "plus or minus". But why 1.96? It represents the middle 95% of the sampling distribution.

- Standard error=SD of sampling distribution= $\sqrt{\hat{p}(1-\hat{p})/n}$, where n =sample size
- How can a survey of 1000 people be used to predict the winner of a presidential election? Assume it's a close race, with the incumbent expecting 52% of the votes. Let's say, by chance, we get a \hat{p} of .53. For $\hat{p}=.53$, SD of sampling distribution = $\sqrt{(.53*.47/1000)}=.016$. Multiplying $.016 * 1.96$, like we did for the other confidence intervals=.03 (or 3%). That's why we often hear the margin of error is 3%.
- We then predict (with 95% confidence) that the incumbent will get between 50 and 56% of the votes. Show calculation.

For elections, because of the electoral college, it makes sense to conduct the survey within each state. Sample size needs would be per state.

Likewise, if data are available by race or other relevant background variables, separate standard deviations and sample sizes can be obtained. **See the cigarette smoking data on the last page.**

For a good discussion of how to interpret poll results, see Myth and Reality in Reporting Sampling Error: How the Media Confuse and

Mislead Readers and Viewers¹.

Hypothesis testing

Hypothesis testing is another way of using what we now know about random sampling distributions to make inferences to the population based on what we know about one sample.

Examples of hypotheses that would be assessed using inferential statistics

- Is there an increase in the proportion of males born in the U.S.?
- Does taking Vioxx increase one's risk of myocardial infarction?
- Is poverty associated with health problems?

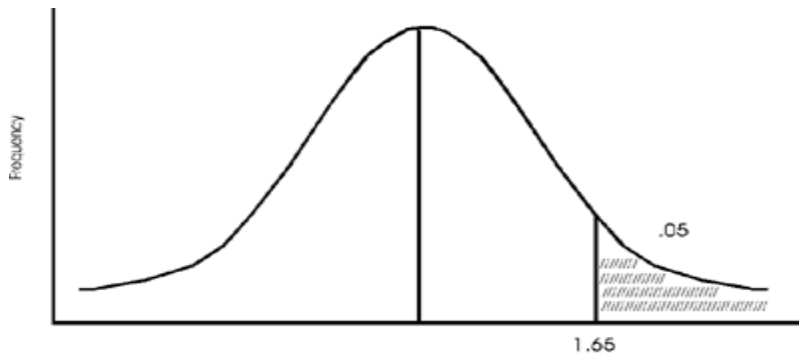
We use the same random sampling distribution as above to see if the sample results are plausible if our hypothesis is correct. For example, **"Reduced Ratio of Male to Female Births in Several Industrial Countries: A Sentinel Health Indicator?"**²

Logic of hypothesis testing:

- My hypothesis is that there is no effect (i.e., proportion of females isn't increasing, taking Vioxx doesn't increase one's risk of myocardial infarction.).
- If there is no effect, 95% of samples will fall in the unshaded region in the figure below. Five % will fall in the outer area. (Note there's one rejection region, while confidence intervals used middle 95% of curve.)
- If we get a sample that falls in the shaded area, we conclude that there is a "statistically significant effect."
- We need to acknowledge that our sample could have been from a population where there was no effect and we just happened to get one of the most extreme 5 percent of the samples.

¹ Taylor, H. (*The Polling Report*, May 4, 1998). Available online at <http://www.pollingreport.com/sampling.htm>

² Davis, DL, Gottlieb, MB, and Stampnitzky, JR (April 1, 1998). [JAMA: The Journal of the American Medical Association](#). 279(13):1018-1023.



(Figure accessed online from <http://www.westgard.com/lesson38.htm>)

Any sample proportions in the unshaded area reflect samples that would not have suggested there's an increase in the proportion of females. Any sample that fell in the tail above 1.65 would suggest that the proportion of females is increasing.

Concluding caveats about statistical significance

Statistical significance doesn't imply real significance. With a large enough sample you can almost be assured of "statistically significant" results. For example, if a chemotherapy treatment which cost \$250,000 produced a significantly significant increase in life expectancy of 15 days, would that be worth giving to other people.

Furthermore, there is always the possibility that you have a sample that is producing a "false positive." (i.e., appear to have detected a significant relationship when you've actually gotten one of the less-likely samples).

On the other hand, lack of statistical significance doesn't mean the relationship doesn't exist. It may mean it wasn't detected. (False negative)

There seems to be a movement away from hypothesis testing toward confidence intervals and p-values (which are described below). This is occurring for many reasons, including the fact that with a large enough sample size you can almost always find significant differences.

"Do not be overly impressed by the complex details of [inference and probability], This elaborate machinery cannot remedy basic flaws in

producing the data such as voluntary response samples [convenience sample] and confounded experiments.” [M&M: p. 416]

(SKIP) P-values.

The p-value is the probability of observing the sample value given the hypothesis of no difference. This is now used more frequently, as an alternative to hypothesis testing. This enables the reader to decide if they want to consider a probability less than .05 (the traditional cut-off for “significance”) or use more or less stringent criteria, particularly in light of multiple comparisons (there are 19 in the table below).

Table showing 95 percent CI and p-value

TABLE 2. Number and percentage of households with one or more persons reporting illness or injury within 1 week after Tropical Storm Allison, by flood status of home — Houston, Texas, June 16, 2001

Condition	Flooded (n=137)		Nonflooded (n=283)		OR*	(95% CI) [†]	p value [‡]
	No.	(%)	No.	(%)			
Illness	35	(25.5)	19	(6.7)	4.7	(1.8– 12.0)	<0.001
Diarrhea/Stomach condition	15	(10.9)	9	(3.2)	6.2	(1.4– 28.0)	0.017
Respiratory symptoms/Cold	14	(10.2)	7	(2.5)	3.2	(0.9– 10.9)	0.046
Headache/Dizziness	10	(7.3)	4	(1.4)	4.4	(0.8– 25.6)	0.056
Anxiety/Distress	5	(3.6)	0	(0.0)	undefined	undefined	0.059
Heart attack/Heart problems	4	(2.9)	0	(0.0)	undefined	undefined	0.059
Chronic illness made worse	3	(2.2)	0	(0.0)	undefined	undefined	0.134
Undefined generalized illness	1	(0.7)	1	(0.4)	undefined	undefined	0.149
Sleep disturbance/Nightmare	12	(8.8)	2	(7.1)	3.3	(0.5– 22.3)	0.240
Rash	2	(1.5)	2	(0.7)	6.0	(0.2–149.6)	0.286
Allergies	0	(0.0)	1	(0.4)	undefined	undefined	0.527
Injury	11	(8.0)	6	(2.1)	1.9	(0.4– 8.4)	0.463
Fall	2	(1.5)	0	(0.0)	undefined	undefined	0.153
Blunt injury	1	(0.7)	0	(0.0)	undefined	undefined	0.387
Insect bite	3	(2.2)	0	(0.0)	undefined	undefined	0.394
Abrasion/Cut/Puncture	2	(1.5)	3	(1.1)	0.4	(0.0– 8.1)	0.596
Auto accident	0	(0.0)	1	(0.4)	undefined	undefined	0.683
Other undefined injury	1	(0.7)	0	(0.0)	undefined	undefined	0.683
Animal bite	2	(1.5)	2	(0.7)	1.0	(0.1– 20.0)	1.000

* Odds ratio.

[†] Confidence interval.

[‡] Analysis of odds ratio, confidence interval, and p value stratified by census tract.

Centers for Disease Control. MMWR Morbidity and Mortality Weekly Report. 2002 May 3;51(17):365-9.

Conclusions

Summarize inferential statistics

Applicability of what we talked about today to other situations, although we won't discuss any of these today.

- **Even if the underlying distribution isn't normal, the sampling distribution can be close enough to normal to be able to use it. This is the powerful part of statistics.**
- Sampling distribution of a mean, F-statistic (ANOVA—3 or more groups), chi-square (non-numeric variable with more than two outcomes), correlation (relationship between 2 variables).

Shameless self-promotion

- I will be presenting at the Virginia Libraries Association on Thursday, October 20th. The session is "Statistics for Librarians." I'll cover designing a survey, trying to get people to respond to your survey, and presenting the results. I won't be discussing inference.
- I will also be giving the Inferential Stats part of today's workshop at SLA in Baltimore this June.

I hope I've given you a sense of what inferential statistics is about. I've described a process where you take sample data and generalize it to a population. We then used a random sampling distribution to compute a confidence interval or describe a result as "statistically significant." The next time you hear the term "confidence interval or statistical significance I hope you'll have a better idea of what the researcher is talking about.

TABLE 1. Prevalence of current smoking* among women aged 18-44 years -- United States, National Health Interview Survey, + 1987-1992³

Characteristic	1987 (n=13,809)		1988 (n=13,746)		1989 (n=6502)		1990 (n=12,954)		1991 (n=13,439)		1992 (n=3717)	
	%	(95% CI &)	%	(95% CI)	%	(95% CI)	%	(95% CI)	%	(95% CI)	%	(95% CI)
Race (Age group {yrs})												
White												
18-24	27.8	(+/-2.2)	27.5	(+/-2.1)	26.0	(+/-3.0)	25.4	(+/-2.2)	25.2	(+/-2.1)	27.2	(+/-4.2)
25-34	31.8	(+/-1.5)	31.0	(+/-1.5)	30.9	(+/-2.3)	28.5	(+/-1.5)	28.4	(+/-1.5)	30.0	(+/-3.0)
35-44	29.2	(+/-1.5)	28.3	(+/-1.5)	26.2	(+/-2.3)	25.0	(+/-1.5)	26.8	(+/-1.5)	27.9	(+/-2.8)
Total	30.0	(+/-1.0)	29.2	(+/-1.0)	28.1	(+/-1.5)	26.5	(+/-1.0)	27.1	(+/-1.0)	28.6	(+/-1.9)
Black												
18-24	20.4	(+/-4.4)	21.8	(+/-4.1)	18.0	(+/-5.5)	10.0	(+/-2.8)	11.9	(+/-3.2)	5.9	(+/-4.2)
25-34	35.8	(+/-3.4)	37.2	(+/-3.6)	28.8	(+/-4.8)	29.1	(+/-3.3)	32.5	(+/-3.6)	29.0	(+/-6.9)
35-44	35.3	(+/-4.3)	27.6	(+/-3.8)	31.4	(+/-5.3)	25.5	(+/-3.6)	35.5	(+/-4.0)	27.9	(+/-7.3)
Total	31.2	(+/-2.5)	30.0	(+/-2.3)	26.6	(+/-3.3)	22.8	(+/-2.1)	28.1	(+/-2.4)	22.6	(+/-4.1)
Ethnicity												
Hispanic	20.0	(+/-2.7)	20.4	(+/-2.5)	21.9	(+/-4.1)	16.9	(+/-2.6)	16.5	(+/-2.1)	18.9	(+/-4.2)
Non-Hispanic	30.6	(+/-1.0)	29.7	(+/-0.9)	28.1	(+/-1.4)	26.6	(+/-1.0)	27.9	(+/-1.0)	27.8	(+/-1.8)
Education (yrs)												
<12	46.5	(+/-2.7)	45.9	(+/-2.7)	42.7	(+/-3.9)	40.6	(+/-2.9)	40.5	(+/-2.7)	40.2	(+/-4.8)
12	33.7	(+/-1.4)	32.7	(+/-1.4)	31.2	(+/-2.1)	31.1	(+/-1.5)	32.0	(+/-1.5)	31.9	(+/-3.0)
13-15	24.7	(+/-1.6)	24.7	(+/-1.6)	25.9	(+/-2.5)	20.6	(+/-1.5)	22.8	(+/-1.7)	24.0	(+/-3.1)
>=16	14.2	(+/-1.5)	13.9	(+/-1.4)	12.0	(+/-2.0)	10.5	(+/-1.3)	12.0	(+/-1.4)	12.5	(+/-2.4)
Socioeconomic status @												
At/Above poverty level	28.3	(+/-1.0)	27.2	(+/-0.9)	26.4	(+/-1.4)	23.6	(+/-0.9)	25.3	(+/-0.9)	24.7	(+/-1.9)
Below poverty level	37.0	(+/-3.1)	38.0	(+/-2.7)	34.9	(+/-3.9)	36.1	(+/-3.1)	32.7	(+/-3.0)	40.0	(+/-4.9)
Unknown	31.1	(+/-4.0)	31.9	(+/-4.2)	28.9	(+/-5.2)	30.4	(+/-3.8)	31.0	(+/-3.3)	24.7	(+/-5.6)
Total	29.6	(+/-0.9)	28.8	(+/-0.9)	27.6	(+/-1.3)	25.6	(+/-0.9)	26.7	(+/-0.9)	26.9	(+/-1.7)

* Smoked at least 100 cigarettes and currently smoking. This analysis excludes persons with unknown smoking status.

+ Health topic supplements: Cancer Control and Epidemiology, 1987; Occupational Health, 1988; Diabetes Risk Factors, 1989; Health Promotion and Disease Prevention, 1990 and 1991; and Cancer Control, 1992.

& Confidence interval.

@ Poverty statistics are based on a definition originated by the Social Security Administration in 1964, subsequently modified by federal interagency committees in 1969 and 1980, and prescribed by the Office of Management and Budget as the standard to be used by federal agencies for statistical purposes.

³ From "Health Objectives for the Nation Cigarette Smoking Among Women of Reproductive Age -- United States, 1987-1992." MMWR Weekly, November 04, 1994 / 43(43);789-791,797. Accessed online from <http://www.cdc.gov/mmwr/preview/mmwrhtml/00033226.htm>

Random Sampling distribution principles

- **Even if the underlying distribution isn't normal, the sampling distribution can be close enough to normal to be able to use it.**
- Assume each sample is exactly the same size.
- Assume you take samples over and over a zillion times.
- Assume each of the samples is chosen at random (any sample has an equal probability of being selected).
- These samples will usually differ slightly. The value of the statistics you compute (e.g., the proportion of males in the sample) will vary from sample to sample.
- The mean of the sampling distribution will equal the population value.
- The standard deviation of the sampling distribution is a function of the proportion and the size of the sample you're working with.
 - The bigger the sample, the narrower the sampling distribution.

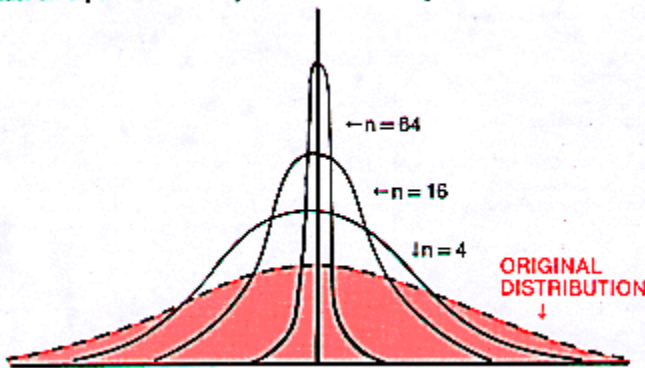


Figure accessed online from http://www.uth.tmc.edu/uth_orgs/educ_dev/oser/FIG3_1.GIF

Sampling distribution of a proportion (non-numeric variable)

Proportion=number of people with the attribute/number of people total.

Proportion of people who voted for Bush

Proportion of librarians who are male

Proportion of people here today who live in Virginia

You could take samples over and over or you could use statistical theory to help you determine the mean and SD of the sampling distribution of the proportion.

(See Figure below.; From M&M, p. 375)

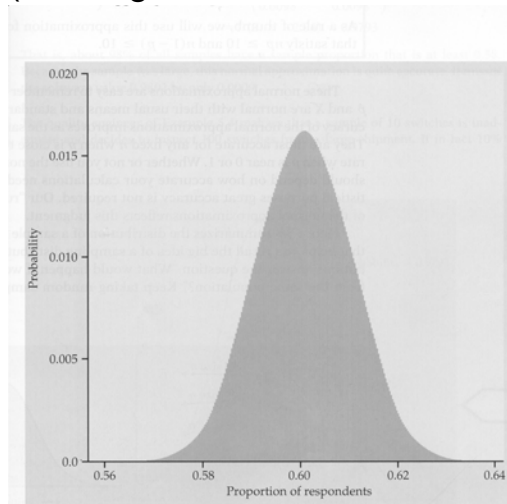


FIGURE 5.2 Probability histogram of the sample proportion \hat{p} based on a binomial count with $n = 2500$ and $p = 0.6$. The distribution is very close to normal.

- Mean of a sampling distribution of a proportion=mean of population
- Standard error=SD of sampling distribution= $\sqrt{\hat{p} (1 - \hat{p})/n}$, where n =sample size
- For $\hat{p} = .6$, SD of sampling distribution = $\sqrt{(.6 * .4/2500)} = .01$
- Sample size affects RSDM. The larger the sample, the more likely the sample proportion will approach the population proportion.
- This figure indicates that if you have a population proportion of .6 and you take samples of size 2500, almost all the samples will have a proportion within .58 and .62 (i.e., very close to the population value). With samples of 25, the distribution will be much more spread out.

Proportion of blue baby shoes from repeated random samplings of 10 shoes (i.e., n=10). See graph on next page.

Sample #

1	.5
2	.6
3	.4
4	.5
5	.1
6	.5
7	.4
8	.7
9	.6
10	.5
rows deleted	
999,998	.0
999,999	.5
1,000,000	.5