5. Statistical Inference: Estimation

Goal: How can we use sample data to estimate values of population parameters?

Point estimate: A single statistic value that is the "best guess" for the parameter value

Interval estimate: An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. Called a ***confidence interval***.

(Based on sampling distribution of the point estimate)

Point Estimators – Most common to use sample values

- Sample mean estimates population mean μ

$$\hat{\mu} = \bar{y} = \frac{\sum y_i}{n}$$

- Sample std. dev. estimates population std. dev. σ

$$\hat{\sigma} = s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

- Sample proportion $\hat{\pi}$ estimates population proportion π

Properties of good estimators

- **Unbiased**: Sampling distribution of the estimator centers around the parameter value

ex. Biased estimator: sample range. It cannot be larger than population range.

- **Efficient**: Smallest possible standard error, compared to other estimators

Ex. If population is symmetric and approximately normal in shape, sample mean is more efficient than sample median in estimating the population mean and median. (can check this with sampling distribution applet at www.prenhall.com/agresti)

Confidence Intervals

- A **confidence interval** (CI) is an interval of numbers believed to contain the parameter value.

- The probability the method produces an interval that contains the parameter is called the **confidence level.** It is common to use a number close to 1, such as 0.95 or 0.99.

- Most CIs have the form

point estimate ± margin of error

with margin of error based on spread of sampling distribution of the point estimator;

e.g., margin of error $\cong$ 2(standard error) for 95% confidence.

Confidence Interval for a Proportion (in a particular category)

- Recall that the sample proportion $\hat{\pi}$ is a mean when we let y = 1 for observation in category of interest, y = 0 otherwise

- Recall the population proportion is mean μ of prob. dist having

$$P(1) = \pi \text{ and } P(0) = 1 - \pi$$

- The standard deviation of this probability distribution is

$$\sigma = \sqrt{\pi(1-\pi)} \text{ (e.g., } 0.50 \text{ when } \pi = 0.50)$$

- The standard error of the sample proportion is

$$\sigma_{\hat{\pi}} = \sigma / \sqrt{n} = \sqrt{\pi(1-\pi)/n}$$

- Recall the sampling distribution of a sample proportion for large random samples is approximately normal

  (Central Limit Theorem)

- So, with probability 0.95, sample proportion $\hat{\pi}$ falls within 1.96 standard errors of population proportion π

- 0.95 probability that

$$\hat{\pi} \text{ falls between } \pi - 1.96\sigma_{\hat{\pi}} \text{ and } \pi + 1.96\sigma_{\hat{\pi}}$$

- Once sample selected, we're 95% confident

$$\hat{\pi} - 1.96\sigma_{\hat{\pi}} \text{ to } \hat{\pi} + 1.96\sigma_{\hat{\pi}} \text{ contains } \pi$$

This is the CI for the population proportion π (almost)

Finding a CI in practice

- Complication: The true standard error

$$\sigma_{\hat{\pi}} = \sigma / \sqrt{n} = \sqrt{\pi(1-\pi)/n}$$

itself depends on the unknown parameter!

In practice, we estimate

$$\sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ by } se = \sqrt{\frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n}}$$

and then find the 95% CI using the formula

$$\hat{\pi} - 1.96(se) \text{ to } \hat{\pi} + 1.96(se)$$

Example: What percentage of 18-22 year-old Americans report being "very happy"?

2006 GSS data: 35 of n = 164 say they are "very happy"
(others report being "pretty happy" or "not too happy")

$$\hat{\pi} = 35/164 = .213 \quad (.31 \text{ for all ages}),$$

$$se = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.213(0.787)/164} = 0.032$$

95% CI is 0.213 ± 1.96(0.032), or 0.213 ± 0.063,
(i.e., "margin of error" = 0.063)

which gives (0.15, 0.28). We're 95% confident the population proportion who are "very happy" is between 0.15 and 0.28.

Find a 99% CI with these data

- 0.99 central probability, 0.01 in two tails
- 0.005 in each tail
- z-score is 2.58
- 99% CI is 0.213 ± 2.58(0.032),

or 0.213 ± 0.083, which gives (0.13, 0.30)

Greater confidence requires wider CI

Recall 95% CI was (0.15, 0.28)

Suppose sample proportion of 0.213 based on n = 656 (instead of 164)

$$se = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.213(0.787)/656} = 0.016 \text{ (instead of } 0.032)$$

95% CI is 0.213 ± 1.96(0.016), or 0.213 ± 0.031, which is (0.18, 0.24).

Recall 95% CI with n = 164 was (0.15, 0.28)

Greater sample size gives narrower CI
(quadruple n to halve width of CI)

These *se* formulas treat population size as infinite (see Exercise 4.57 for *finite population correction*)

Some comments about CIs

- Effects of *n*, confidence coefficient true for CIs for other parameters also
- If we repeatedly took random samples of some fixed size *n* and each time calculated a 95% CI, in the long run about 95% of the CI's would contain the population proportion π.

  (CI applet at www.prenhall.com/agresti)
- The probability that the CI does *not* contain π is called the **error probability**, and is denoted by α.
- α = 1 – confidence coefficient

$(1-\alpha)100\%$	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.050	1.645
95%	.05	.025	1.96
99%	.01	.005	2.58

- General formula for CI for proportion is

$$\hat{\pi} \pm z(se) \text{ with } se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

z-value such that prob. for a normal dist within *z* standard errors of mean equals confidence level

(e.g., *z*=1.96 for 95% confidence, *z*=2.58 for 99% conf.)

- With *n* for most polls (roughly 1000), margin of error usually about ± 0.03 (ideally)
- Method requires "large *n*" so sampling distribution of sample proportion is approximately normal (CLT) and estimate of true standard error is decent

  In practice, ok if at least 15 observ. in each category

(ex.: *n*=164, 35 "very happy", 164 – 35 = 129 not "very happy", condition satisfied)

- Otherwise, sampling distribution is skewed

  (can check this with sampling distribution applet at www.prenhall.com/agresti, e.g., for n = 30 but π = 0.1 or 0.9)

  and sample proportion may then be poor estimate of π, and *se* may then be a poor estimate of true standard error.

Example: Estimating proportion of vegetarians (p. 129)

n = 20, 0 vegetarians, sample proportion = 0/20 = 0.0,

$$se = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.0(1.0)/20} = 0.000$$

95% CI for population proportion is 0.0 ± 1.96(0.0), or (0.0, 0.0)

Better CI method (due to Edwin Wilson at Harvard in 1927, but not in most statistics books):

Do not estimate standard error, but figure out π values for which

$$|\hat{\pi} - \pi| \;=\; 1.96\sqrt{\pi(1-\pi)/n}$$

Example: for *n* = 20 with $\hat{\pi} = 0$,

solving the quadratic equation this gives for π provides solutions 0 and 0.16, so 95% CI is (0, 0.16)

- Agresti and Coull (1998) suggested using ordinary CI, estimate ± *z(se),* after adding 2 observations of each type. This simpler approach works well even for very small *n*

  (95% CI has same midpoint as Wilson CI);

Example*:* 0 vegetarians, 20 non-veg; change to 2 veg, 22 non-veg, and then we find

$$\hat{\pi} = 2/24 = 0.083,\; se = \sqrt{(0.083)(0.917)/24} = 0.056$$

95% CI is 0.08 ± 1.96(0.056) = 0.08 ± 0.11, gives (0.0, 0.19).

Confidence Interval for the Mean

- In large random samples, the sample mean has approximately a normal sampling distribution with mean μ and standard error $\sigma_{\bar{y}} = \sigma / \sqrt{n}$
- Thus,

$$P(\mu - 1.96\sigma_{\bar{y}} \leq \bar{y} \leq \mu + 1.96\sigma_{\bar{y}}) = .95$$

- We can be 95% confident that the sample mean lies within 1.96 standard errors of the (unknown) population mean

- Problem: Standard error is unknown (σ is also a parameter). It is estimated by replacing σ with its point estimate from the sample data:

$$se = \frac{s}{\sqrt{n}}$$

95% confidence interval for μ :

$$\bar{y} \pm 1.96(se), \text{ which is } \bar{y} \pm 1.96\frac{s}{\sqrt{n}}$$

This works ok for "large *n*," because *s* then a good estimate of σ (and CLT applies). But for small *n*, replacing σ by its estimate *s* introduces extra error, and CI is not quite wide enough unless we replace *z*-score by a slightly larger "*t*-score."

The *t* distribution (*Student's t*)

- Bell-shaped, symmetric about 0
- Standard deviation a bit larger than 1 (slightly thicker tails than standard normal distribution, which has mean = 0, standard deviation = 1)
- Precise shape depends on **degrees of freedom** (*df*). For inference about mean,

  $df = n - 1$

- Gets narrower and more closely resembles standard normal distribution as *df* increases

  (nearly identical when $df > 30$)

- CI for mean has margin of error $t(se)$,

  (instead of z(se) as in CI for proportion)

Part of a *t* table

	Confidence Level			
	90%	95%	98%	99%
df	t.050	t.025	t.010	t.005
1	6.314	12.706	31.821	63.657
10	1.812	2.228	2.764	3.169
30	1.697	2.042	2.457	2.750
100	1.660	1.984	2.364	2.626
infinity	1.645	1.960	2.326	2.576

df = ∞ corresponds to standard normal distribution

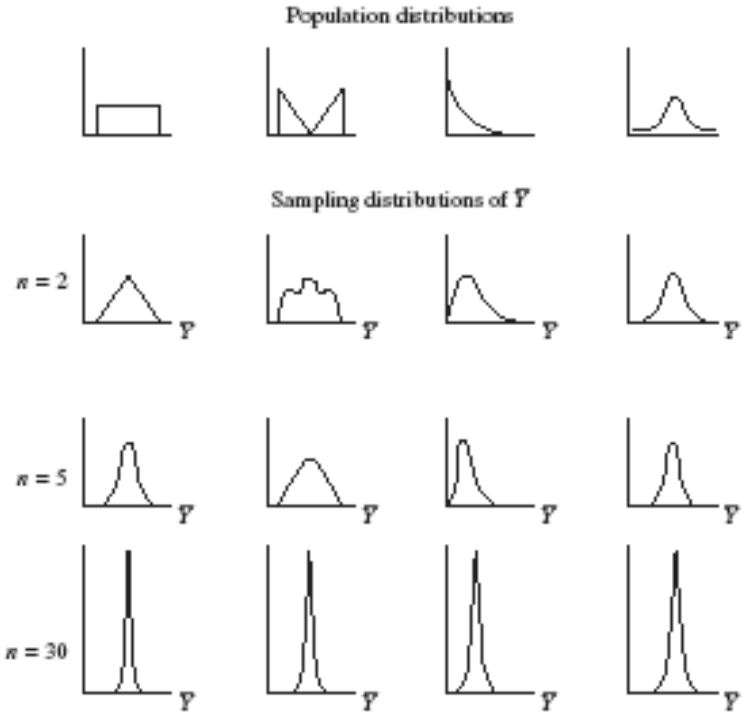
CI for a population mean

- For a random sample *from a normal population distribution*, a 95% CI for μ is

$$\bar{y} \pm t_{.025}(se), \text{ with } se = s/\sqrt{n}$$

where *df* = *n* - *1* for the *t*-score

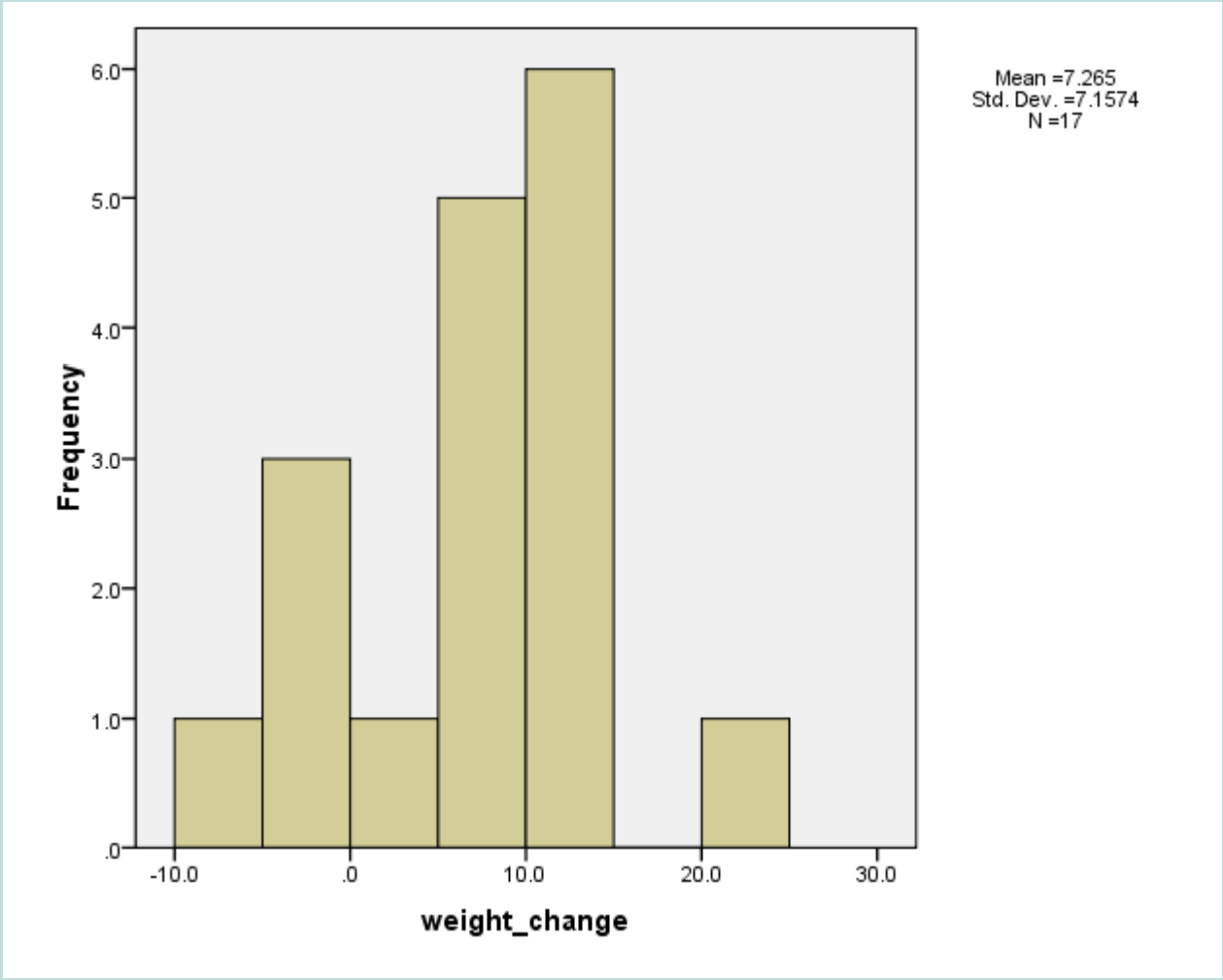
- Normal population assumption ensures sampling distribution has bell shape for *any n* (Recall figure on p. 93 of text and next page). More about this assumption later.


Population distributions
Sampling distributions of $\bar{Y}$
$n = 2$
$n = 5$
$n = 30$
$\bar{Y}$

Example: Anorexia study (p.120)

- Weight measured before and after period of treatment
- *y* = weight at end – weight at beginning
- Example on p.120 shows results for "cognitive behavioral" therapy. For *n*=17 girls receiving "family therapy" (p. 396),

y = 11.4, 11.0, 5.5, 9.4, 13.6, -2.9, -0.1, 7.4, 21.5, -5.3, -3.8, 13.4, 13.1, 9.0, 3.9, 5.7, 10.7


Mean =7.265
Std. Dev. =7.1574
N =17
Frequency
6.0
5.0
4.0
3.0
2.0
1.0
.0
-10.0
.0
10.0
20.0
30.0
weight_change

Software reports

Variable	N	Mean	Std.Dev.	Std. Error Mean
weight_change	17	7.265	7.157	1.736

se obtained as

$$se = s/\sqrt{n} = 7.157/\sqrt{17} = 1.736$$

Since n = 17, df = 16, t-score for 95% confidence is 2.12

95% CI for population mean weight change is

$$\bar{y} \pm t(se), \text{ which is } 7.265 \pm 2.12(1.736), \text{ or } (3.6, 10.9)$$

We can predict that the population mean weight change μ is positive (i.e., the treatment is effective, on average), with value of μ between about 4 and 11 pounds.

Example: TV watching in U.S.

GSS asks "On average day, how many hours do you personally watch TV?"

For n = 899, $\bar{y}$ = 2.865, s = 2.617

What is a 95% CI for population mean?

df = *n-1* = 898 huge, so *t*-score essentially same as z = 1.96

Show *se* = 0.0873, CI is 2.865 $\pm$ 0.171, or (2.69, 3.04)

Interpretation?

Multiple choice:

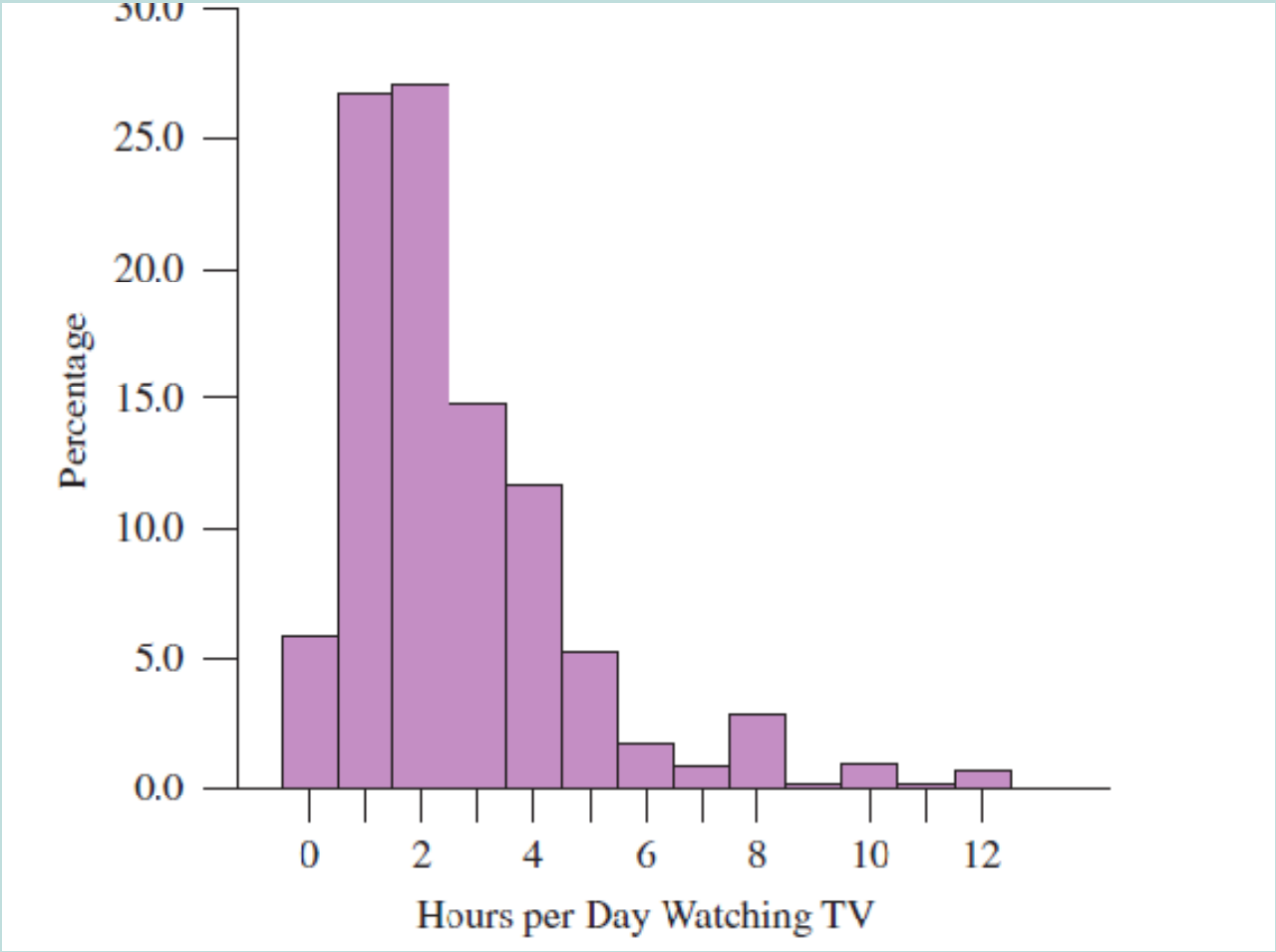
a. We can be 95% confident the sample mean is between 2.69 and 3.04.

b. 95% of the population watches between 2.69 and 3.04 hours of TV per day

c. We can be 95% confident the population mean is between 2.69 and 3.04

d. If random samples of size 899 were repeatedly selected, in the long run 95% of them would contain $\bar{y}$ = 2.865

Note: The *t* method for CIs assumes a *normal* population distribution. Do you think that holds here?


30.0
25.0
20.0
15.0
10.0
5.0
0.0
Percentage
0
2
4
6
8
10
12
Hours per Day Watching TV

Comments about CI for population mean μ

- The method is ***robust*** to violations of the assumption of a normal population dist.

(But, be careful if sample data dist is very highly skewed, or if severe outliers. Look at the data.)

- Greater confidence requires wider CI

- Greater *n* produces narrower CI

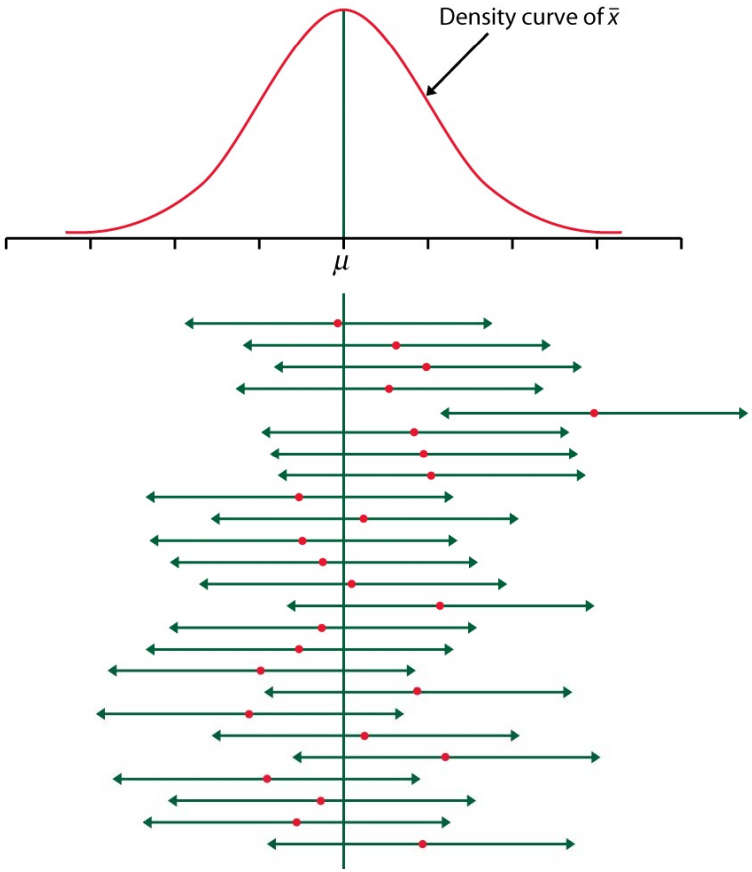
- *t* methods developed by the statistician William Gosset of Guinness Breweries, Dublin (1908)

Because of company policy forbidding the publication of company work in one's own name, Gosset used the pseudonym *Student* in articles he wrote about his discoveries (sometimes called "Student's *t* distribution")

He was given only small samples of brew to test (why?), and realized he could not use normal *z*-scores after substituting *s* in standard error formula.

In the "long run," 95% of 95% CIs for a population mean μ will actually cover μ

(In graph, each line shows a CI for a particular sample with its own sample mean, taken from the *sampling distribution* curve of possible sample means)



Choosing the Sample Size

Ex. How large a sample size do we need to estimate a population proportion (e.g., “very happy”) to within 0.03, with probability 0.95?

i.e., what is *n* so that margin of error of 95% confidence interval is 0.03?

Set 0.03 = margin of error and solve for *n*

$$0.03 = 1.96\sigma_{\hat{\pi}} = 1.96\sqrt{\pi(1-\pi)/n}$$

Solution

$$n = \pi(1-\pi)(1.96/0.03)^2 = 4268\pi(1-\pi)$$

Largest *n* value occurs for π = 0.50, so we'll be "safe" by selecting *n* = 4268(0.50)(0.50) = 1067.

If only need margin of error 0.06, require

$$n = \pi(1-\pi)(1.96/0.06)^2 = 1067\pi(1-\pi)$$

(To double precision, need to quadruple *n*)

What if we can make an educated "guess" about proportion value?

- If previous study suggests population proportion is roughly about 0.20, then to get margin of error 0.03 for 95% CI,

$$n = \pi(1-\pi)(1.96/0.03)^2 = 4268\pi(1-\pi)$$

$$= 4268(0.20)(0.80) = 683$$

- It's "easier" to estimate a population proportion as the value gets closer to 0 or 1 (close election difficult)

- Better to use approx value for π rather than 0.50 unless you have no idea about its value

Choosing the Sample Size

- Determine parameter of interest (population mean or population proportion)
- Select a margin of error (*M*) and a confidence level (determines *z*-score)

Proportion (to be "safe," set $\pi = 0.50$):

$$n = \pi(1-\pi)\left(\frac{z}{M}\right)^2$$

Mean (need a guess for value of σ):

$$n = \sigma^2\left(\frac{z}{M}\right)^2$$

Example: *n* for estimating mean

Future anorexia study: We want *n* to estimate population mean weight change to within 2 pounds, with probability 0.95.

- Based on past study, guess σ = 7

$$n = \sigma^2 \left(\frac{z}{M} \right)^2 = 7^2 \left(\frac{1.96}{2} \right)^2 = 47$$

Note: Don't worry about memorizing formulas such as for sample size. Formula sheet given on exams.

Some comments about CIs and sample size

- We've seen that *n* depends on confidence level (higher confidence requires larger *n*) and the population variability (more variability requires larger *n*)

- In practice, determining *n* not so easy, because (1) many parameters to estimate, (2) resources may be limited and we may need to compromise

- CI's can be formed for any parameter.

  (e.g., see pp. 130-131 for CI for median)

Using *n-1* (instead of *n*) in *s* reduces bias in the estimation of the population std. dev. σ

- **Example**: A binary probability distribution with $n = 2$

y	$P(y)$
0	½
2	½

$\mu = \sum yP(y) = 1, \quad \sigma^2 = \sum (y-\mu)^2 P(y) = 1$ so $\sigma = 1$

Possible samples (equally likely)	$\frac{\Sigma(y_i - \bar{y})^2}{n}$	$\frac{\Sigma(y_i - \bar{y})^2}{n-1}$	$\frac{\Sigma(y_i - \mu)^2}{n}$
(0, 0)	0	0	1
(0, 2)	1	2	1
(2, 0)	1	2	1
(2, 2)	0	0	1
Mean of estimates	0.5	1.0	1.0

- *Confidence interval* methods were developed in the 1930s by Jerzy Neyman (U. California, Berkeley) and Egon Pearson (University College, London)

- The *point estimation* method mainly used today, developed by Ronald Fisher (UK) in the 1920s, is **maximum likelihood**. The estimate is the value of the parameter for which the observed data would have had greater chance of occurring than if the parameter equaled any other number.

  (picture)

- The **bootstrap** is a modern method (Brad Efron) for generating CIs without using mathematical methods to derive a sampling distribution that assumes a particular population distribution. It is based on repeatedly taking samples of size *n (with replacement)* from the sample data distribution.

Using CI Inference in Practice (or in HW)

- What is the variable of interest?
  quantitative – inference about *mean*
  categorical – inference about *proportion*

- Are conditions satisfied?
  - Randomization (why? Needed so sampling dist. and its standard error are as advertised)
  - Other conditions?
    Mean: Look at data to ensure distribution of data not such that mean is irrelevant or misleading parameter
    Proportion: Need at least 15 observ's in category and not in category, or else use different formula (e.g., Wilson's approach, or first add 2 observations to each category)