we skip 8.2 for now

8.3 Confidence Intervals for Proportions

Recall:
- How many cigarettes do you smoke?
  (a means problem) μ, σ or $\bar{x}, s$
  census / sample
- Do you smoke?
  (a proportions problem) $\hat{p} = \frac{x}{N}$

In 8.1 we were building a confidence Intrvl for means

- Here we build conf. Intrvl for proportions

For means we used grouped data

$$SD = \sigma_{pop} / \sqrt{n}$$

we replaced $z = \frac{x-\mu}{\sigma}$ with $z = \frac{x-\mu}{SD}$

- Here we use

$$SE = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- p = pop. proportion
- $\hat{p}$ = sample proportion
- $q = 1 - p$
- $\hat{q} = 1 - \hat{p}$

All the steps for forming a conf. Intvl are the same.

→ The **Assumptions** BUT (aka Conditions) change.

→ The **SE changes** AND but the procedure is the same.

same formula

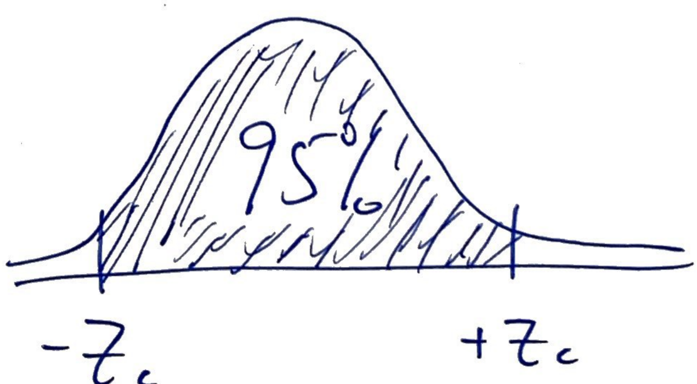
same critical values

- $ME = z_c \cdot SE$ ← use $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ now

-OR-

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

Recall z_c is the z-score that separates the middle 90%, or 95%, etc sample data from the extreme data found in the wings



Finally we use

- CI: $\hat{p} \pm ME$

The **Conditions** are

1. Independent individuals (same)

→ SRS and $n < 10\%$ of population

2. Need sufficiently large data in the sample

→ Success & failures need to be equal to or larger than 10.

$\{ \underbrace{n \cdot p}_{\text{successes}} \overset{?}{\geq} 10 \ \& \ \underbrace{n \cdot (1-\hat{p})}_{\text{failures}} \overset{?}{\geq} 10 \}$

Proportions (z) — Assumptions / Justification

Using the Chart...

- **One sample**
  1. Individuals are independent. — 1. SRS and $n < 10\%$ of the population.
  2. Sample is sufficiently large. — 2. Successes and failures each ≥ 10.
- **Two Groups**
  1. Groups are independent. — 1. (Think about how the data were collected.)
  2. Data in each group are independent. — 2. Both are SRSs and $n < 10\%$ of populations OR random allocation.
  3. Both samples are sufficiently large. — 3. Successes and failures each ≥ 10 for both groups.

Means (t)

- **One Sample** ($df = n - 1$)
  1. Individuals are independent. — 1. SRS and $n < 10\%$ of the population.
  2. Population has a Normal model. — 2. Histogram is unimodal and symmetric.*
- **Matched pairs** ($df = n - 1$)
  1. Data are matched. — 1. (Think about the design.)
  2. Individuals are independent. — 2. SRS and $n < 10\%$ OR random allocation.
  3. Population of differences is Normal. — 3. Histogram of differences is unimodal and symmetric.* or n>30
- **Two independent samples** (df from technology)
  1. Groups are independent. — 1. (Think about the design.)
  2. Data in each group are independent. — 2. SRSs and $n < 10\%$ OR random allocation.
  3. Both populations are Normal. — 3. Both histograms are unimodal and symmetric.* or both n>30

Distributions/Association (χ^2)

- **Goodness of fit** (df = # of cells − 1; one variable, one sample compared with population model)
  1. Data are counts. — 1. (Are they?)
  2. Data in sample are independent. — 2. SRS and $n < 10\%$ of the population.
  3. Sample is sufficiently large. — 3. All expected counts ≥ 5.
- **Homogeneity** [$df = (r - 1)(c - 1)$; many groups compared on one variable]
  1. Data are counts. — 1. (Are they?)
  2. Data in groups are independent. — 2. SRSs and $n < 10\%$ OR random allocation.
  3. Groups are sufficiently large. — 3. All expected counts ≥ 5.
- **Independence** [$df = (r - 1)(c - 1)$; sample from one population classified on two variables]
  1. Data are counts. — 1. (Are they?)
  2. Data are independent. — 2. SRSs and $n < 10\%$ of the population.
  3. Sample is sufficiently large. — 3. All expected counts ≥ 5.

Regression (t, $df = n - 2$)

- **Association of each quantitative variable** ($\beta = 0$?)
  1. Form of relationship is linear. — 1. Scatterplot looks approximately linear.
  2. Errors are independent. — 2. No apparent pattern in residuals plot.
  3. Variability of errors is constant. — 3. Residuals plot has consistent spread.
  4. Errors have a Normal model. — 4. Histogram of residuals is approximately unimodal and symmetric, or Normal probability plot reasonably straight.*

(*less critical as n increases)

t-table

We will introduce(d) this table in 8.2 (means w/no σ)

TABLE A-3 t Distribution: Critical t Values

			Area in One Tail		
	0.005	0.01	0.025	0.05	0.10
			Area in Two Tails	Confidence Intervals	
Degrees of Freedom	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372
11	3.106	2.718	2.201	1.796	1.363
12	3.055	2.681	2.179	1.782	1.356
13	3.012	2.650	2.160	1.771	1.350
14	2.977	2.624	2.145	1.761	1.345
15	2.947	2.602	2.131	1.753	1.341
16	2.921	2.583	2.120	1.746	1.337
17	2.898	2.567	2.110	1.740	1.333
18	2.878	2.552	2.101	1.734	1.330
19	2.861	2.539	2.093	1.729	1.328
20	2.845	2.528	2.086	1.725	1.325
21	2.831	2.518	2.080	1.721	1.323
22	2.819	2.508	2.074	1.717	1.321
23	2.807	2.500	2.069	1.714	1.319
24	2.797	2.492	2.064	1.711	1.318
25	2.787	2.485	2.060	1.708	1.316
26	2.779	2.479	2.056	1.706	1.315
27	2.771	2.473	2.052	1.703	1.314
28	2.763	2.467	2.048	1.701	1.313
29	2.756	2.462	2.045	1.699	1.311
30	2.750	2.457	2.042	1.697	1.310
31	2.744	2.453	2.040	1.696	1.309
32	2.738	2.449	2.037	1.694	1.309
34	2.728	2.441	2.032	1.691	1.307
36	2.719	2.434	2.028	1.688	1.306
38	2.712	2.429	2.024	1.686	1.304
40	2.704	2.423	2.021	1.684	1.303
45	2.690	2.412	2.014	1.679	1.301
50	2.678	2.403	2.009	1.676	1.299
55	2.668	2.396	2.004	1.673	1.297
60	2.660	2.390	2.000	1.671	1.296
65	2.654	2.385	1.997	1.669	1.295
70	2.648	2.381	1.994	1.667	1.294
75	2.643	2.377	1.992	1.665	1.293
80	2.639	2.374	1.990	1.664	1.292
90	2.632	2.368	1.987	1.662	1.291
100	2.626	2.364	1.984	1.660	1.290
200	2.601	2.345	1.972	1.653	1.286
300	2.592	2.339	1.968	1.650	1.284
400	2.588	2.336	1.966	1.649	1.284
500	2.586	2.334	1.965	1.648	1.283
750	2.582	2.331	1.963	1.647	1.283
1000	2.581	2.330	1.962	1.646	1.282
2000	2.578	2.328	1.961	1.646	1.282
Large	2.576	2.326	1.960	1.645	1.282
CL	99%	98%	95%	90%	80%

Z_c values are found conviently on the t-table

Norm Dist → Large; ← Z_c

* For proportions we do not use D.O.F or this t-table
* It is just a convenient one table stop

4

EX A sample of 800 parents in the HART School district found that 632 parents favor music education. Construct a 90% confidence int'vl for the proportion of parents favoring music education

---

Step 0: Type of problem: 1 pop proportion
$\Rightarrow$ z-table

Assumptions:

- Independence? is 800 parents more than 10% of all parents in the Hart School Dist? I.E. are there 8000 families in the Hart School dist. Ans: most probably (Assume so).

((too large of sample?))

- Large enough sample?

((large enough sample size?))

  - Successes : $632 \overset{?}{\geq} 10$, yes
  - Failures : $\underset{\text{success}}{800-632} = \underset{\text{fails}}{168} \overset{?}{\geq} 10$, yes

Step 1: Point estimate

successes

$$\hat{p} = \frac{x}{n} = \frac{632}{800} = 0.79$$

or 79%

Step 2: (a) Conf. Level: 90% $(\alpha = 1 - 0.9 = 0.10)$

(b) critical value: $z_c = 1.645$

Step 3: Standard Error

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

$$= \sqrt{\frac{0.79(1-0.79)}{800}} = 0.0144$$

Step 4: Margin of Error: $z_c \cdot SE$

$$ME = (1.645)(0.0144) = 0.0237$$

Step 5: C. Intvl: $\hat{p} \pm ME$

$\hat{p}$ − ME … $\hat{p}$ + ME

$$0.79 - 0.0237 < p < 0.79 + 0.0237$$

$$0.766 < p < 0.814$$

Step 6: We are 90% confident that the proportion of HSD parents who favor music education is between 76.6% and 81.4%

(5)

MATH 140 Statistics **Confidence Interval Worksheet** Name Santa Claus

STEP 0: (a) Type of problem (circle the line or part therein)

- (1- pop) | 2 pop for proportion (z-table)
- 1- pop | 2 pop for means (t-table)

(b) Assumptions (state the general and justify your application's)

- Independence? — There are more than 10*800 = 8000 families in the Hart School District. ✓
- Large enough Sample? — 632 ≥ 10 & (800-632) ≥ 10 ✓

STEP 1: Compute the point estimate

$\hat{p} = 632/800 = 0.79$

STEP 2: (a) State the Confidence Level: 90%

(b) Find the corresponding critical value from the tables (reverse z-look up)

row 1.60 column 0.4 and 0.5

critical value (circle one): (z) t = 1.645

STEP 3: Compute the standard error.

(a) Formula SE: $\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$ $\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$ $\frac{s}{\sqrt{n}}$ $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ $\frac{s_d}{\sqrt{n}}$

$1-0.79=0.21$

$SE = \sqrt{\frac{(0.79)(0.21)}{800}} = 0.0144$

SE Value = 0.0144

STEP 4: Compute the Margin of Error = critical value * SE

ME = 1.645 * 0.0144 = 0.0237

STEP 5: Construct the Confidence Interval: point estimate ± ME

$0.79 - 0.0237 < p < 0.79 + 0.0237$

$0.766 < p < 0.814$

STEP 6: Interpret the results

We are 90% confident that the proportion of Hart District parents who believe music education is benficial is between 76.6% and 81.4%

(*) Necc'y Sample Size

In order to acheive the desired confidence Level we can compute "n" needed

- proportions – Recall

$$ME = \pm Z_c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Solve this for "n"

$$n = \hat{p}(1-\hat{p})\left(\frac{Z_c}{ME}\right)^2$$

($\hat{p}$ ← point estimate from a preliminary sample)

Ex

Q: Do you smoke? y/n

STEPS
1. Take initial sample
2. calculate $\hat{p} = \#\text{ yes}/\text{TOT }\#$
3. get Z_c from the tables & Decide on how close of interval you want {±0.5}
4. Calculate "n" from the formula above.
5. Do you need more samples?

IF you can't obtain an initial sample use $\hat{p} = 0.50$, then $\hat{q} = 0.50$ also

$$n = 0.25\left(\frac{Z_c}{m}\right)^2$$

This is the worst case scenario But a good start

An example on next page ⇒

Ex A sample of 800 parents found that 632 believe music education to be important.

Q: What is the necc'y sample size to acheive a 95% confidence level with a margin of error $\pm$ 0.025 i.e. 0.25%

---

1) 635 of 800
2) $\hat{p} = 635/800 = \underline{0.79}$
3) 95% → $z = \underline{1.96}$

4) $n = (0.79)(0.21)\left(\frac{1.16}{\pm 0.025}\right)^2$ from the formula

$n = 1019.71$ round up always **1020**

5) So, to be able to acheive 95% with a ME of $\pm$ 0.025 on the proportion, we need an additional 1020 − 800 = **220 samples**

$\hat{p} \pm 0.025$ pretty accurate, no?

The more acc'y we desire the more samples are needed.

WARNING: we may run up against the independent sample criteria of more than 10% being polled