11.2 Hyp. Test for the difference between Two proportions

Notation

p_1 & p_2 are pop. proportions

$\hat{p}_1$ & $\hat{p}_2$ are sample proportions

x_1 & x_2 are number of success in each sample respectly

n_1 & n_2 are sample sizes

Assumptions

1. both samples must be S.R.S.
2. each population must be 20 times larger, or more, than the sample.
3. There are two categories in each sample
4. each sample has 10 successes & 10 fails.

Test-Statistic

- pooled proportion $= \dfrac{x_1 + x_2}{n_1 + n_2} = \hat{p}$

combined successes ÷ combined samples

- $SE = \sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$

alternatively

$SE = \sqrt{\dfrac{\hat{p}\hat{q}}{n_1} + \dfrac{\hat{p}\hat{q}}{n_2}}$

$\hat{q} = 1 - \hat{p}$

- $\boxed{z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{SE}}$

null: no diff.

11.2 HT tests for difference between proportions

Proportions (z)

- **One sample**
  1. Individuals are independent.
  2. Sample is sufficiently large.

  1. SRS and $n < 10\%$ of the population.
  2. Successes and failures each ≥ 10.

z critical value

- **Two Groups** Assumptions
  1. Groups are independent.
  2. Data in each group are independent.
  3. Both samples are sufficiently large.

  Justifications
  1. (Think about how the data were collected.)
  2. Both are SRSs and $n < 10\%$ of populations OR random allocation.
  3. Successes and failures each ≥ 10 for both groups.

  each population is 20 times as large as its sample

Means (t)

- **One Sample** ($df = n - 1$)
  1. Individuals are independent.
  2. Population has a Normal model.

  1. SRS and $n < 10\%$ of the population.
  2. Histogram is unimodal and symmetric.*

- **Matched pairs** ($df = n - 1$)
  1. Data are matched.
  2. Individuals are independent.
  3. Population of differences is Normal.

  1. (Think about the design.)
  2. SRS and $n < 10\%$ OR random allocation.
  3. Histogram of differences is unimodal and symmetric.* or n>30

- **Two independent samples** (df from technology)
  1. Groups are independent.
  2. Data in each group are independent.
  3. Both populations are Normal.

  1. (Think about the design.)
  2. SRSs and $n < 10\%$ OR random allocation.
  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.
  2. Data in sample are independent.
  3. Sample is sufficiently large.

  1. (Are they?)
  2. SRS and $n < 10\%$ of the population.
  3. All expected counts ≥ 5.

- **Homogeneity** [$df = (r - 1)(c - 1)$; many groups compared on one variable]
  1. Data are counts.
  2. Data in groups are independent.
  3. Groups are sufficiently large.

  1. (Are they?)
  2. SRSs and $n < 10\%$ OR random allocation.
  3. All expected counts ≥ 5.

- **Independence** [$df = (r - 1)(c - 1)$; sample from one population classified on two variables]
  1. Data are counts.
  2. Data are independent.
  3. Sample is sufficiently large.

  1. (Are they?)
  2. SRSs and $n < 10\%$ of the population.
  3. All expected counts ≥ 5.

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

- **Association of each quantitative variable** ($\beta = 0$?)
  1. Form of relationship is linear.
  2. Errors are independent.
  3. Variability of errors is constant.
  4. Errors have a Normal model.

  1. Scatterplot looks approximately linear.
  2. No apparent pattern in residuals plot.
  3. Residuals plot has consistent spread.
  4. Histogram of residuals is approximately unimodal and symmetric, or Normal probability plot reasonably straight.*

(*less critical as n increases)

Hypothesis

$H_0 : p_1 = p_2$ or $p_1 - p_2 = 0$

$H_A : p_1 \neq p_2$ or $p_1 < p_2$ or $p_1 > p_2$

two tail — left tail — right tail

Ex Are older, more experienced workers, less likely to use computers at work vs. younger workers?

Gen. Social Survey shows • 259 of 350 workers in the $\underbrace{25-40}_{\text{pop }1}$ yr. use computers to do their work

• 384 of 500 workers in the $\underbrace{41-65}_{\text{pop }2}$ yr used computer to do their work.

use $\alpha = 0.05$

Test to see if younger workers use computers more frequently than older workers.

preliminary observations

$\hat{p}_1 = \frac{259}{350} = 0.74$ young

$\hat{p}_2 = \frac{384}{500} = 0.768$ old

$\hat{p}_1 - \hat{p}_2$ is $\underline{\underline{-0.028}}$

So the survey actually shows the older gen is using compt. more.

Do younger workers use computers more than older workers?

Data from our survey

Sample	size	use a computer	proportion
① 25-40 yo	$n_1 = 350$	$x_1 = 259$	$\hat{p}_1 = 259/350 = 0.740$
② 41-65 yo	$n_2 = 500$	$x_2 = 384$	$\hat{p}_2 = 384/500 = 0.768$

STEP 0: (a) 2-pop proportions

(b)
- SRS each ⟶ assumed
- $N_1 > 20 n_1$ ⟶ if we expand scope
- $N_2 > 20 n_2$ ⟶ if we expand scope
- Two categories ⟶ "use" vs "not use" ✓
- 10 successes ⟶ $259 > 10$; $384 > 10$ ✓
- 10 failures ⟶ $350 - 259 > 10$; $500 - 384 > 10$ ✓

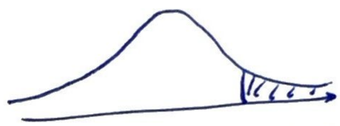
(c) $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{259 + 384}{350 + 500} = \underline{\underline{0.7565}}$ {pooled proportion

STEP 1: (a) $H_0: p_1 = p_2$

$H_A: p_1 > p_2$

young old

(b)



(c) Right-Tail Test

critical value method

STEP 2: $\alpha = 0.05$ given to us for the test

STEP 3: (a) $SE = \sqrt{0.7565(1 - 0.7565)\left[\frac{1}{350} + \frac{1}{500}\right]} = \underline{\underline{0.0299}}$

(b) $Z_{test} = \dfrac{(0.740 - 0.768) - 0}{0.0299} = \boxed{-0.94}$

STEP 4: $Z_{crit} = 1.645$ {see attached t-table but we read z}

STEP 5: The test statistic is not further away than the critical value.

STEP 6: We cannot conclude that the younger generation uses computers more often in the workplace.

HT 2pop - proportions (z-table)

TABLE A-3	*t* Distribution: Critical *t* Values				
	our example is an "Area in One Tail" test case with				
	0.005	0.01	0.025	α = 0.05	0.10
Degrees of Freedom	Area in Two Tails				
	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	$\boxed{1.645}$	1.282

Z_{crit}

Step 7: Statdisk

$\boxed{\text{Analysis} \rightarrow \text{hyp testing} \rightarrow \text{proportion two samples}}$

- select $p_1 > p_2$
- Keep $\alpha = 0.05$ {significance}
- Sample 1 $n_1 = 350$
  $x_1 = 259$
- Sample 2 $n_2 = 500$
  $x_2 = 384$

$\boxed{\text{Evaluate}}$

Results

$H_A: p1 > p2$ (right tail)

$z_{\text{test stat}} = \boxed{-0.93604}$

p-value $= \boxed{0.825}$, this is

(very large since $p_1 < p_2$ to start with and yet we were asked to test $p_1 > p_2$, nevertheless)

such a sample could still result