Stat 145, Wed 31-Mar-2021 -- Wed 31-Mar-2021
Biostatistics
Spring 2021

Wednesday, March 31st 2021

Wk 9, We
Topic:: Inference on two proportions
Read:: Lock5 6.7-6.9

0. Wetsuits
1. 379 of 460 females support tougher gun-control laws, 318 of 520 males
2. 10 of 24 cocaine addicts treated with desipramine had relapses, compared with 20 of 24 who received placebo

Have discussed 2-proportion CI construction

Summary of that

Use $\hat{p}_1 - \hat{p}_2$ as point estimate

Use $SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Obtain your critical value from norm(0, 1) (call it z^*)

Need to discuss: hypothesis testing for difference of two proportions

$H_0: p_1 - p_2 = 0$, $\qquad H_a: p_1 - p_2 \neq 0$

Standardize test statistic: $\quad z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{\hat{p}_1 - \hat{p}_2}}$

New wrinkle: How SE is approximate for hyp. tests is different than for CI (2-proportion settings)

1-proportion settings

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

CI: Insert $\hat{p}$ for p

Hyp tests: Insert null value for p

2-proportion settings

$$SE_{\hat{p}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

CI: Take $\hat{p}_1$ for p_1
$\hat{p}_2$ for p_2

hyp. tests: $\dfrac{\text{total of successes}}{n_1 + n_2}$

Ex.)

24 cocaine addicts given drug (desipramine), 10 relapsed

24 " " " placebo, 20 relapsed

$H_0: p_P - p_D = 0$, $H_a: p_P - p_D \neq 0$

test stat (unstandardized): $\hat{p}_P - \hat{p}_D = \frac{20}{24} - \frac{10}{24} = \frac{10}{24}$

Pool results together $\tilde{p} = \frac{10+20}{24+24} = \frac{30}{48}$ prop. of relapsers

Use this pooled proportion in my SE

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n_1} + \frac{\tilde{p}(1-\tilde{p})}{n_2}}$$

$$= \sqrt{\frac{30/48\,(1-30/48)}{24} + \frac{30/48\,(1-30/48)}{24}}$$

To standardize my test stat

$$z = \frac{(\hat{p}_p - \hat{p}_o) - 0}{SE}$$

Class notes from STAT 145

Thomas Scofield

March 31, 2021

Paired (quantitative) data

We carry out an analysis of **Wetsuits** data set, containing paired data from swimming times both with and without wetsuits.

We intend to carry out an hypothesis test on the difference of times `Wetsuit - NoWetsuit`, with hypotheses

$$\mathbf{H}_0 : \mu_D = 0, \;\; \mathbf{H}_a : \mu_D \neq 0.$$

```
mutate(Wetsuits, difference = Wetsuit - NoWetsuit) -> myDat
favstats(~difference, data=myDat)
```

```
##   min     Q1 median  Q3  max   mean         sd  n missing
##  0.05 0.0575   0.08 0.1 0.11 0.0775 0.02179449 12       0
```

My standardized test statistic is:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{0.0775}{0.02179/\sqrt{12}} = 12.318.$$

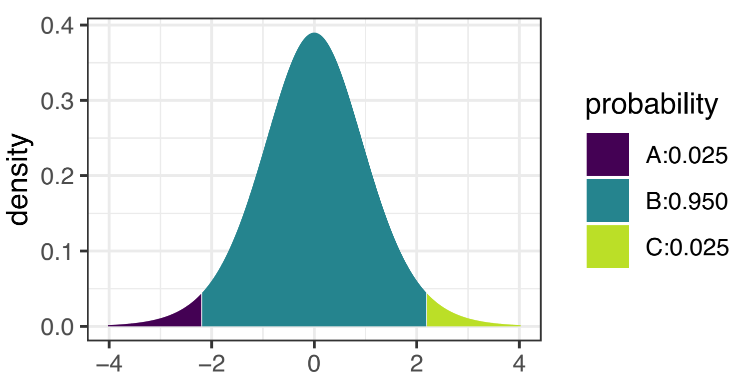
Say we intend to use significance level $\alpha = 0.05$. My null distribution is a t-distribution with $df = 11$, so this significance level establishes tails with cutoffs

```
qt( c(.025, .975), df=11 )
```

```
## [1] -2.200985  2.200985
```

There is a command, `xpt()`, I use below to make a StatKey-style picture of this t-distribution with its corresponding tails:

```
xpt(c(-2.2, 2.2), df=11)
```




```
## [1] 0.02504306 0.97495694
```

The rejection region is comprised of the two tails, colored differently. Since our standardized test statistic falls in the upper tail, we will reject $\mathbf{H}_0$ in favor of the alternative. To say more precisely what our P-value is, we use the `pt()` command:

```
(1 - pt(12.318, df=11))*2
```

```
## [1] 8.886537e-08
```

Example of 2-proportion hypothesis testing

An experiment for treating cocaine addicts to help prevent relapse involved giving desipramine or placebo and measuring relapse rates (proportions).

The hypotheses:

$$\mathbf{H}_0 : p_D - p_P = 0 \text{ vs. } \mathbf{H}_a : p_D - p_P \neq 0.$$

The sample data showed

- 14 of 24 addicts receiving desipramine did *not* relapse
- 4 of 24 addicts receiving placebo did *not* relapse

Our (nonstandardized) test statistic is

$$\widehat{p}_D - \widehat{p}_P = \frac{14}{24} - \frac{4}{24} = \frac{10}{24} \doteq 0.41667.$$

Since we are doing an hypothesis test, here, we calculate the **pooled proportion** $\tilde{p}$:

$$\tilde{p} = \frac{14+4}{24+24} = \frac{18}{48} = 0.375.$$

This is the number we use in place of *both* p_1 and p_2 in our approximate standard error. That is,

$$\mathrm{SE}_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

becomes

$$\mathrm{SE}_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n_1} + \frac{\tilde{p}(1-\tilde{p})}{n_2}} = \sqrt{\tilde{p}(1-\tilde{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

This means we will estimate $\mathrm{SE}_{\widehat{p}_1 - \widehat{p}_2}$ to be

$$\mathrm{SE} \approx \sqrt{(0.375)(1-0.375)\left(\frac{1}{24} + \frac{1}{24}\right)} \doteq 0.13975.$$

Using this, we standardize our test statistic (calling it z):

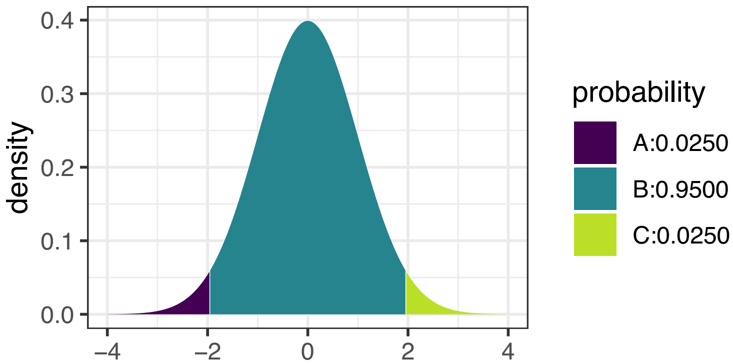
$$z = \frac{(\text{unstandardized}) - (\text{null value})}{\mathrm{SE}} = \frac{0.41667}{0.13975} \doteq 2.9814.$$

Our null distribution for the standardized test statistic is Norm(0,1). If we take $\alpha = 0.05$, then the rejection region is displayed below as before, now using the `xpnorm()` command (not a command you need to know, but it allows me to insert pictures like those you see in the StatKey normal calculator):

```
xpnorm(c(-1.96, 1.96), mean=0, sd=1)
```

```
##
```

```
## If X ~ N(0, 1), then
##  P(X <= -1.96) = P(Z <= -1.96) = 0.025   P(X <=  1.96) = P(Z <=  1.96) = 0.975
##  P(X >  -1.96) = P(Z >  -1.96) = 0.975   P(X >   1.96) = P(Z >   1.96) = 0.025
##
```




```
## [1] 0.0249979 0.9750021
```

Since our standardized test statistic $z = 2.98$ is in the rejection region (the upper tail, specifically), we will reject $\mathbf{H}_0$ in favor of $\mathbf{H}_a$. We can obtain the actual P-value using `pnorm()`:

```
2 * (1-pnorm(2.9814))
```

```
## [1] 0.002869337
```