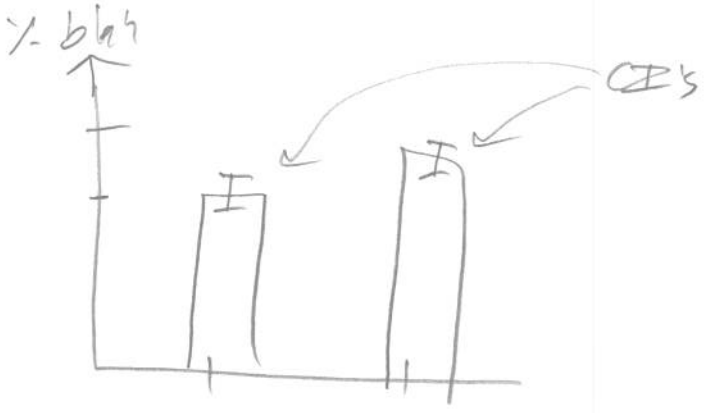
6/23/11

Lecture #17

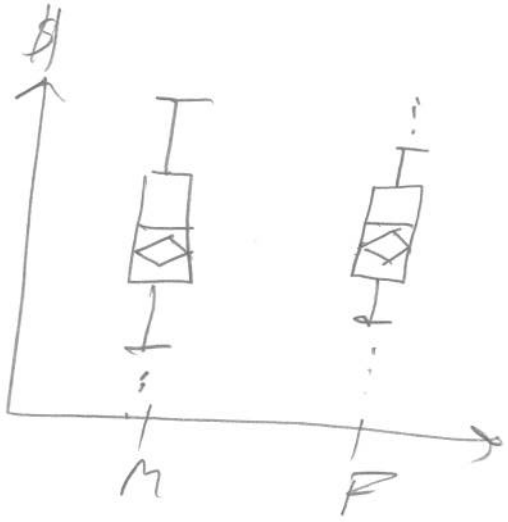
Plan

- 2-prop & 2-samp hyp. testing

Yesterday we talked about bivariate analysis of categorical vs. interval data:

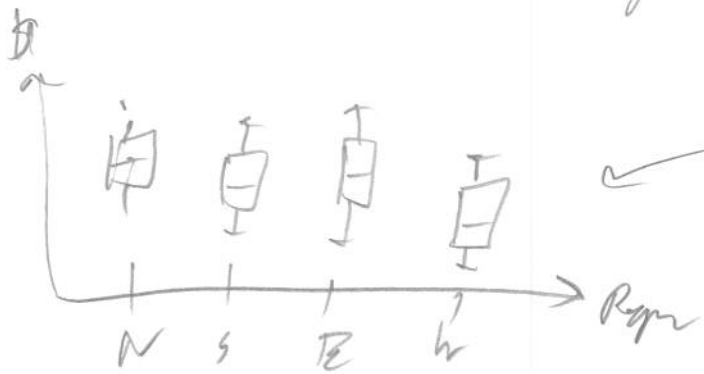


(proportions)



Histogram

In this class, the categorical variable will only have two states ⇒ "two groups" which are different



comparisons such as these are called ANOVA (Stat 102, 112)

The normal thing to do is to test!

1) If the proportions between the two groups are the same (i.e. $p_1 = p_2$), 2-prop z-test

2) If the means between the two groups are the same (i.e. $(\mu_1 = \mu_2)$, 2-smple t-test

Or... to build CI's for the difference in the proportions or means.

Some of 2-prop tests and CI's 2-parameters

$$\hat{p}_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right), \quad \hat{p}_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

We can observe the difference:

$$D = \hat{p}_1 - \hat{p}_2$$

← distr of diffs

→ of course for these to be true we need $n\hat{p}_1 > 10$, $n(p_1) > 10$, $n\hat{p}_2 > 10$, $n(1-\hat{p}_2) > 10$

the difference we get in an experiment is $d = \hat{p}_1 - \hat{p}_2$ a draw from $D = \hat{p}_1 - \hat{p}_2$

What is D's distribution? Here's substituting the formula:

$$SE[\hat{p}_1 - \hat{p}_2] = \sqrt{Var[\hat{p}_1 - \hat{p}_2]}$$

Assume both groups are i.i.d.

$$= \sqrt{Var\,\hat{p}_1 + Var\,\hat{p}_2}$$

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

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Now we do an experiment. We get d, we standardize:

$$Z = \frac{D - E[D]}{SE[D]}, \quad z = \frac{d - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \quad \text{(exact)}$$

Let's see an example: p#37

35 of 60 customers in the East prefer red fabric ~~of~~ 32 of 72 customers in the West prefer red fabric. Is there any real difference in red fabric preferences or difference due to chance? Test at $\alpha = 5\%$

Let p_1 be the % of customers in the E; Let p_2 be the % of customers in the West

H_0: $p_1 - p_2 = 0$ i.e. no difference

H_a: $p_1 - p_2 \neq 0$ i.e. there is a difference (two-tailed)

$\hat{p}_1 = \frac{35}{60} = .5833$, $\hat{p}_2 = \frac{32}{72} = .4444$

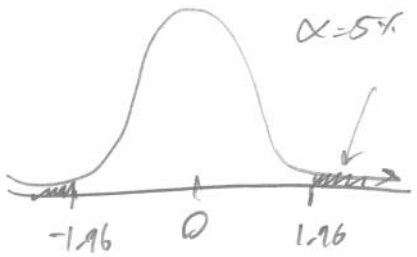
Let's test it... ~~Find~~ compute z statistic first:

$$z = \frac{d - (0)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{.5833 - .4444}{\sqrt{\frac{.5833(1-.5833)}{60} + \frac{.4444(1-.4444)}{72}}}$$

don't know p_1, p_2 just diff!

$$= \frac{.1389}{0.08645} = 1.607 < z_{0.025} = 1.96$$

$\Rightarrow$ fail to reject



$$p_{val} = P(|Z| \geq 1.607) = 2 \cdot .054 = .108$$

$> \alpha = 0.05$

$\Rightarrow$ fail to reject

Compute a 95% CI for the difference:

$$CI_{p_1 - p_2, 95\%} = \left[d \pm z_{0.025} \cdot SE(D) \right]$$

$$\approx \left[(\hat{p}_1 - \hat{p}_2) \pm z_{0.025} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

$$= \left[.1389 \pm 1.96 \cdot 0.08645 \right]$$

$$= \left[-0.031, 0.308 \right)$$

Since $0 \in CI_{p_1 - p_2, 95\%} \Rightarrow$ fail to reject null hypothesis $p_1 - p_2 = 0$ at $\alpha = 5\%$

You will need to study left-tailed and right-tailed on your own

Now for the more interesting case (goes to final graph)
we are looking for difference in means. 0-params!

$$\bar{X}_1 \sim N\left(\mu_1, \left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2\right), \quad \bar{X}_2 \sim N\left(\mu_2, \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2\right)$$

$$SE\left[\bar{X}_1 - \bar{X}_2\right] = \sqrt{Var\left[\bar{X}_1 - \bar{X}_2\right]} = \sqrt{Var\left[\bar{X}_1\right] + Var\left[\bar{X}_2\right]} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$0 = \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)^2\right) = N\left(\mu_1 - \mu_2, \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)^2\right)$$

if $\sigma_1 = \sigma_2$

$d = \bar{X}_1 - \bar{X}_2$ is a draw from $0 = \bar{X}_1 - \bar{X}_2$

Further complication: What if we don't know σ_1, σ_2?

We only have $s_1, s_2 \Rightarrow T$ distribution $\Rightarrow$ (Need both pop's same)

but what is ~~the~~ SE and what is the df?

We consider two situations:

$$T = \frac{d - E[0]}{SE[0]}$$

① σ_1, σ_2 different: if $\frac{s_1^2}{s_2^2} > 3$ or $\frac{s_2^2}{s_1^2} > 3$

$$SE\left[\bar{X}_1 - \bar{X}_2\right] \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad df = \min\{n_1, n_2\}$$

do not use formula on p. 445
it's too complicated

② $\sigma = \sigma_1 = \sigma_2$ same: if $\frac{s_1^2}{s_2^2} \leq 3$

$$SE\left[\bar{X} - \bar{X}_2\right] \approx \text{spooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{where spooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}, \quad df = n_1 + n_2 - 2$$

Example p429. Website testing between two designs, new, old.
300 visitors: 169 old spent $\bar{x}_{old} = \$253$, $s_{old} = \$130$, 131 new spent $\bar{x}_{new} = \$328$, $s_{new} = \$161$
Let μ_1 be the mean for new website
Let μ_2 be the mean for old website.
histogram passes Q-Q plot for normality

Is the new site better for revenue generation? $\alpha = 5\%$.

$H_0: \mu_1 - \mu_2 = 0$, $H_a: \mu_1 - \mu_2 > 0$ (higher revenue)

~~Which~~ 2-Samp t-test do we do?

$$\frac{s_1^2}{s_2^2} = \frac{\$161^2}{\$130^2} = 1.53 < 3 \Rightarrow \text{pooled test...}$$

$df = n_1 + n_2 - 2 = 298 \overset{>3}{\Rightarrow}$ use z

$$t = \frac{d - (0)}{SE[0]} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\$328 - \$253}{\sqrt{\frac{(131-1)161^2 + (169-1)130^2}{131+169-2}}\sqrt{\frac{1}{131}+\frac{1}{169}}}$$

$$= \frac{75}{199.39 \cdot 0.116} = \frac{75}{16.803} = 4.46 > t_{0.15,298} \approx z_{0.05} = 1.96$$

$\Rightarrow$ reject H_0

$$p\text{-val} = P(T_{298} \geq 4.46) \approx P(Z \geq 4.46) \approx 0 < \alpha = 5\%$$

$\Rightarrow$ reject H_0

The new website is probably better than the old website

Airbag treatments: experimental vs. new airbags crash cars

detection + inflation 70ms. New designs to do it faster...

Crash $n_2 = 4$ cars old system $\bar{X} = 72$ ms, $s = 5$ ms

Crash $n = 5$ cars new system $\bar{X} = 67$ ms, $s = 2.8$ ms. $\alpha = 1\%$

Let μ_1 be the mean of new system

Let μ_2 be the mean of old system

$H_0: \mu_1 - \mu_2 = 0$

$H_a: \mu_1 - \mu_2 < 0$ (left-tailed)

$\frac{s_2^2}{s_1^2} = \frac{5\text{ms}^2}{2.8\text{ms}^2} = 3.2 > 3 \Rightarrow$ unpooled

$t = \frac{d - (0)}{SE(0)} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{67 - 72}{\sqrt{\frac{2.8^2}{5} + \frac{5^2}{4}}} = \frac{-5}{2.79} = -1.78$

$df = \min\{n_1, n_2\} = \min\{4, 5\} = 4$

> -2.132

$\Rightarrow$ fail to reject

$p\text{-val} = P(T_4 \leq -1.78) = .0788 > \alpha = 5\% \Rightarrow$ fail to reject

Ignore "paired comparisons" pp 36–41

Done with categorical vs. numeral bivariate analysis!

read Ch 10 for Wed! and numeral vs. numeral

X, Y independent $P(X = x, Y = y) = P(X = x)\,P(Y = y)$

or cont. $f_{XY}(x, y) = f_X(x)\, f_Y(y)$

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{XY}(x,y)\, dx\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\, dx\, dy$$

$= \int$