Stat 312: Lecture 18
Other two sample tests

Moo K. Chung
mchung@stat.wisc.edu

November 11, 2004

Concepts

1. *Paired data.* For a given paired sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ with $\mathbb{E}X_i = \mu_X$ and $\mathbb{E}Y_i = \mu_Y$, a test statistic for testing $H_0 : \mu_X = \mu_Y$ can be based on one sample test.

2. Let $X_i \sim$ Bernoulli(p_1) and $Y_j \sim$ Bernoulli(p_2). Let $\hat{p}_1 = \sum_{i=1}^n X_i/n$ and $\hat{p}_2 = \sum_{j=1}^m Y_j/m$.
$$\mathbf{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}.$$

3. Suppose p_1 and p_2 denote the proportion of individuals in population 1 and 2 respectively. For sufficiently large n and m, we use a Z-statistic for testing
$$H_0 : p_1 = p_2 \text{ vs. } H_1 : p_1 \neq p_2 :$$
$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \mathbb{E}(\hat{p}_1 - \hat{p}_2)}{\sqrt{\mathbf{Var}(\hat{p}_1 - \hat{p}_2)}}.$$

Examples

Example 1. 10 students took two midterm exams.
Student || 01 02 03 04 05 06 07 08 09 10
Midterm 1|| 80 75 60 90 99 60 55 85 65 70
Midterm 2|| 70 60 70 72 95 66 60 80 70 60
Is the first exam easier than the second exam? Test it at level 0.05.
Solution. Let X_i and Y_i be the first and the second midterm scores for the i-th student. Perform the t-test on $H_0 : \mu_Y - \mu_X = 0$ based on observations $Y_i - X_i$.

```
> x<-c(80, 75, 60, 90, 99, 60, 55,
85, 65, 70)
> y<-c(70, 60, 70, 72, 95, 66, 60,
80, 70, 60)
> t.test(y-x, conf.level=0.95)
 One Sample t-test
data:  y - x t = -1.1739, df = 9,
p-value = 0.2706 alternative
hypothesis: true mean is not equal
to 0 95 percent confidence
interval:
 -10.537282   3.337282
```

Example 2. Is it harmful to use Marijuana when mothers are pregnant?

	User	Nonuser
Sample size	1246	11,178
Birth defects	42	294

Solution. Let p_1 and p_2 be the birth defect proportions among Marijuana users and non users. The hypothesis we are testing is
$$H_0 : p_1 = p_2 \text{ vs. } H_0 : p_1 > p_2.$$
From data $\hat{p}_1 = 42/1246 = 0.034, \hat{p}_2 = 294/11178 = 0.026$. Under $H_0 : p_1 = p_2 = p$, $\hat{p} = (42 + 294)/(1246 + 11,178) = 0.027$. Then the estimate for the variance in Concept 2 under H_0 would be $\hat{p}(1-\hat{p})(1/n + 1/m) = 0.0048^2$. Then $z = 1.53$. We reject H_0 if z is larger so the P-value is $P(Z > 1.53) = 1 - P(Z < 1.53) = 1 - \texttt{pnorm}(1.53) =$ 0.06. So we do not reject H_0 at 0.05 level but we reject H_0 at 0.1 level.

Example 3. There are two coins. You toss the first coin 100 times and observed 40 heads. When you threw the second coin 110 times, you observed 50 heads. Test if the two coins give the same number of heads.
Solution. Let p_1 and p_2 be the probabilities of getting heads for the first and the second coins respectively.
$$H_0 : P_1 = p_2 \text{ vs. } H_1 : p_1 \neq p_2.$$
Let X be the number of heads when you threw the first coin $n = 100$ times and Y be the number of heads when you threw the second coin $m = 110$ times. $\hat{p}_1 - \hat{p}_2 = X/n - Y/m$. Then the test statistic is $Z = \frac{X/n - Y/m}{\sqrt{p(1-p)(1/n+1/m)}} \sim N(0,1)$, where $p = p_1 = p_2$. Since p is unknown, you estimate it by pooling the samples: $\hat{p} = (x+y)/(m+n)$. $\hat{p} = (40 + 50)/(100+110) = 0.43$ and $z = -0.60$. Since we reject H_0 if $|z|$ is large, the P-value would be $P(|Z| > 0.6) = 2P(Z > 0.6) = 2(1 - P(Z \leq 0.6)) = 2(1 - 0.73) = 0.54$. So we do not reject H_0 at any level smaller than 0.5.

Review problems

Example 9.9, 9.11.
Note: Starting lecture 19, we will jump to Chapter 12 and study linear regressions.