Two Sample Tests
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-09-08 13:47

Two Sample Test Settings

- Single cross-sectional sample, comparing two sub-samples
- Compare samples from two different populations (2 cross-sectional samples, case control study)
- Single sample; subjects randomly allocated to different interventions (experiment, clinical trials)

Fundamentals

- Def 5.2 Two rvs Y_1 and Y_2 are *independent* if for all y_1 and y_2

$$\Pr[Y_1 \leq y_1, Y_2 \leq y_2] = \Pr[Y_1 \leq y_1] \Pr[Y_2 \leq y_2]$$

- Result 5.1 If Y_1 and Y_2 are independent rvs, then for any two constants a_1 and a_2 the rv $W = a_1 Y_1 + a_2 Y_2$ has mean and variance

$$E(W) = a_1 E(Y_1) + a_2 E(Y_2)$$

$$V(W) = a_1^2 V(Y_1) + a_2^2 V(Y_2)$$

Fundamentals

- Result 5.2 If Y_1 and Y_2 are independent rvs that are **normally** distributed, then $W = a_1 Y_1 + a_2 Y_2$ is normally distributed with mean and var given by Result 5.1
- Corrollary: If $\bar{Y}_1$ and $\bar{Y}_2$ are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, then

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Fundamentals

- Result 5.3 If $\bar{Y}_1$ and $\bar{Y}_2$ are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and the same variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Note $n_1 = n_2$ implies

$$s_p^2 = \frac{1}{2}(s_1^2 + s_2^2)$$

Two Sample t-test Example

- An experiment was conducted to see if a drug could prevent premature birth
- 30 women at risk of premature birth were randomly assigned to take drug or placebo (15 in each group)
- Endpoint: birthweight

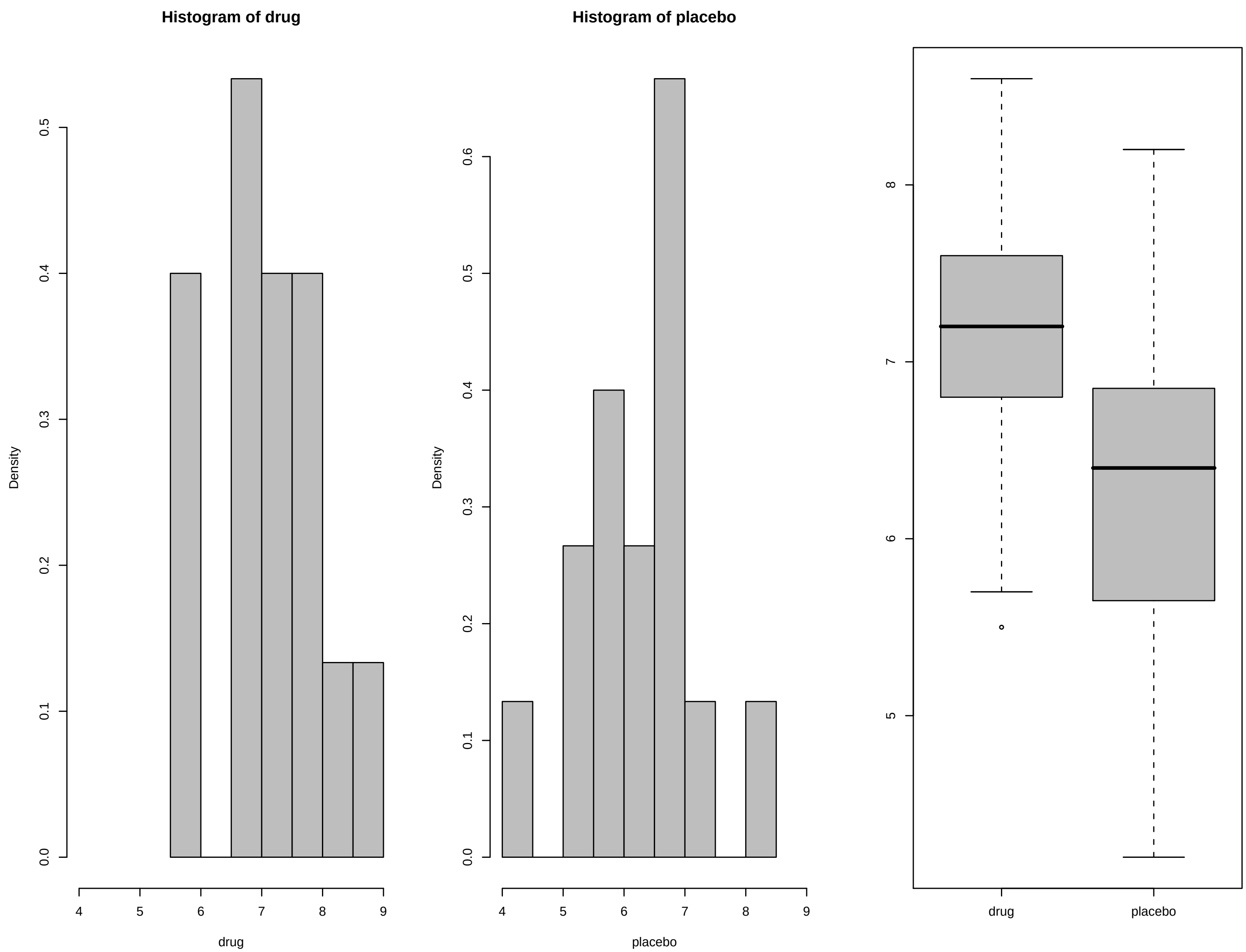
Two Sample t-test Example

- Let 1=drug, 2=placebo
- $H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 > \mu_2$
- $C_\alpha = \{t : t > t_{1-\alpha;28}\}$
- $C_{.05} = \{t : t > 1.7\}$

Two Sample t-test Example

Drug	Placebo
6.9	6.4
7.6	6.7
7.3	5.4
7.6	8.2
6.8	5.3
7.2	6.6
8.0	5.8
5.5	5.7
5.8	6.2
7.3	7.1
8.2	7.0
6.9	6.9
6.8	5.6
5.7	4.2
8.6	6.8

Two Sample t-test Example



Two Sample t-test Example

- $\bar{y}_1 = 7.08 \; s_1 = 0.899$
- $\bar{y}_2 = 6.26 \; s_2 = 0.961$
- Thus

$$s_p^2 = \frac{14(.899)^2 + 14(.961)^2}{28} = 0.8695$$

$$t = \frac{7.08 - 6.26}{.931\sqrt{2/15}} = 2.41$$

- Since $t \in C_{.05}$, reject H_0

$$p = 1 - F_{t_{28}}(2.41) = .011$$

Two Sample t-test: BW example

- R

```
> t.test(bw$drug,bw$placebo,var.equal=TRUE,alternative="greater")

        Two Sample t-test

data:  bw$drug and bw$placebo
t = 2.4136, df = 28, p-value = 0.01129
alternative hypothesis: true difference in means is greater than 0
```

Two Sample t-test: BW example

- SAS

```
proc ttest;  class trt;  var bw;

                     The TTEST Procedure

                            T-Tests

Variable   Method          Variances     DF   t Value   Pr > |t|

bw         Pooled          Equal         28      2.41     0.0226
bw         Satterthwaite   Unequal     27.9      2.41     0.0226
```

Homogeneity of Variance

- Want to test

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ versus } H_A : \sigma_1^2 \neq \sigma_2^2$$

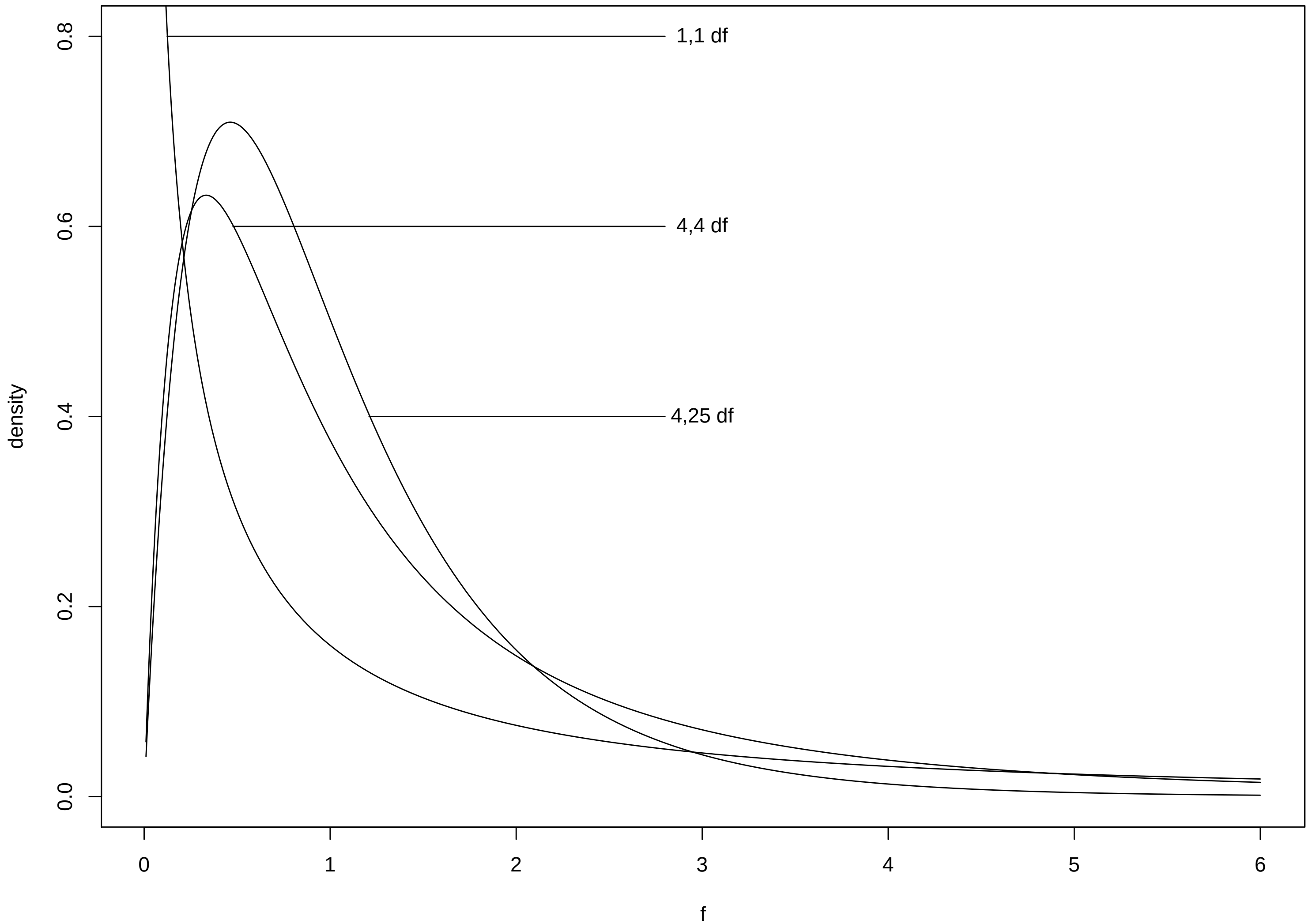
- We know that (assuming normality)

$$\frac{(n_k - 1)s_k^2}{\sigma_k^2} \sim \chi^2_{n_k-1} \text{ for } k = 1, 2$$

- If X_1 and X_2 are independent rvs with $X_1 \sim \chi^2_{v_1}$ and $X_2 \sim \chi^2_{v_2}$, then

$$\frac{X_1/v_1}{X_2/v_2} \sim F_{v_1,v_2}$$

F Distribution



Homogeneity of Variance

- Let
$$X_k = \frac{(n_k - 1)s_k^2}{\sigma_k^2} \text{ for } k = 1, 2$$
- It follows that
$$Y = \frac{X_1/(n_1 - 1)}{X_2/(n_2 - 1)} \sim F_{n_1-1,n_2-1}$$
- Thus
$$Y = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

Homogeneity of Variance

- Under $H_0 : \sigma_1^2 = \sigma_2^2$, such that
$$\frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$
- For $H_A : \sigma_1^2 \neq \sigma_2^2$, reject null if s_1^2/s_2^2 is v large or v small (i.e., near zero)
- Formally,

$$C_\alpha = \{f : f < F_{n_1-1, n_2-1, \alpha/2} \text{ or } f > F_{n_1-1, n_2-1, 1-\alpha/2}\}$$

where $f = s_1^2/s_2^2$

Homogeneity of Variance

- Note: $F_{v_1,v_2,\alpha} = 1/F_{v_2,v_1,1-\alpha}$
- Table A.5 and A.6 of text for two-sided $\alpha = .10$ and $\alpha = .02$; see errata
- R

```
> qf(.975,14,14)
[1] 2.978588
```

- SAS

```
data; y = finv(.975,14,14);
```

Homogeneity of Variance: BW example

- $H_0 : \sigma_1^2 = \sigma_2^2$; $H_A : \sigma_1^2 \neq \sigma_2^2$
- For $\alpha = .05$,

$$C_{.05} = \{f : f < F_{14,14,.025} \text{ or } f > F_{14,14,.975}\}$$

$$= \{f : f < 0.34 \text{ or } f > 2.98\}$$

- Observed test statistic

$$f = \frac{.8994^2}{.9605^2} = 0.8768$$

- Therefore, do not reject H_0

$$p = 2 * F_{14,14}(.8768) = 0.809$$

Homogeneity of Variance: BW example

- SAS

```
proc ttest;  class trt;  var bw;

                     The TTEST Procedure

                            T-Tests

Variable   Method          Variances     DF   t Value   Pr > |t|

bw         Pooled          Equal         28      2.41     0.0226
bw         Satterthwaite   Unequal     27.9      2.41     0.0226


                     Equality of Variances

 Variable    Method       Num DF    Den DF    F Value    Pr > F

 bw          Folded F         14        14       1.14    0.8090
```

Homogeneity of Variance: BW example

- R

```
> var.test(bw$drug,bw$placebo)

        F test to compare two variances

data:  bw$drug and bw$placebo
F = 0.8767, num df = 14, denom df = 14, p-value = 0.809
alternative hypothesis: true ratio of variances is not equal to 1
```

Homogeneity of Variance

- Cf page 133 of text
- Genuine interest in whether vars equal
- WRT testing $H_0 : \mu_1 = \mu_2$
  - For small samples, potential for type II error
  - For large samples, CLT/Slutsky
  - Adjustment for sequential testing
- For additional reading, see Moser and Stevens (TAS 1992)

Testing $\mu_1 = \mu_2$

- What if $\sigma_1^2 \neq \sigma_2^2$ and unknown?
- Solutions

  1. Large sample approximation
  2. Normality: Welch-Satterthwaite approximation (Behrens-Fisher problem)
  3. Transformation
  4. Nonparametric methods: Wilcoxon Ranksum

Large Sample Approximation

- If n_1 and n_2 are large, homoegeneity of variance assumption is not important
- Recall CLT plus Slutsky implies

$$\bar{Y} \sim N(\mu, \frac{s^2}{n})$$

- Thus

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\mu_1 - \mu_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})$$

Large Sample Approximation

- Therefore, to test $H_0 : \mu_1 - \mu_2 = \delta$, we can use
$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- Under H_0, $Z \sim N(0, 1)$
- Approximation gets better as $n_1, n_2 \to \infty$
- Generally, require $n_j \geq 25$ for $j = 1, 2$
- Note assumption that Y's normally distributed no longer needed either (CLT)

Large Sample Approximation: Example

- A study was done to compare the percent body fat of 3^{rd} graders at schools on 2 Native American reservations: Tohona and Apache
- $H_0 : \mu_T = \mu_A$ vs $H_A : \mu_T \neq \mu_A$
- $n_T = 63$, $n_A = 35$
- $C_{.05} = \{z : |z| > 1.96\}$
- $\bar{y}_T = 37.9\%$; $s_T = 8.66$; $\bar{y}_A = 32.8\%$; $s_A = 6.88$

$$z = \frac{37.9 - 32.8}{\sqrt{\frac{8.66^2}{63} + \frac{6.88^2}{35}}} = 3.2; p = 0.0014$$

Welch-Satterthwaite approximation

- Assume normality; n_1, n_2 small
- Statistic

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$$

- Note 5.2 of text:

$$df_{text} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}} - 2$$

Welch-Satterthwaite approximation

- Welch (Biometrika 1938), SAS

$$df_{welch} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

- Use $\lfloor df \rfloor$ if using tables

Welch-Satterthwaite approximation: Example

- Premature birth example

$$n_1 = n_2 = 15$$

$$s_1 = 0.8994, s_2 = 0.9605$$

$$df_{text} = 29.86, df_{welch} = 27.88$$

Welch-Satterthwaite approximation: R

- R

```
> t.test(drug,placebo,var.equal=FALSE)

        Welch Two Sample t-test
data:  drug and placebo
t = 2.4136, df = 27.88, p-value = 0.02262

alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1239229 1.5160771

sample estimates:
mean of x mean of y
     7.08      6.26
```

Summary

Normal	Var known	Var equal	N large	Test statistic
$\surd$	$\surd$			$\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
	$\surd$		$\surd$	
$\surd$		$\surd$		$\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$
			$\surd$	$\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim N(0,1)$
$\surd$				$\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{df}$
				Transform, nonparametrics