Statistical Inference

Applied Regression and Other Multivariable Methods
Sections 3-5 – 3-7

Statistical Inference

- Two general categories of inference
  - **Estimation:**

    Quantifies a specific population parameter

    Point estimate and confidence intervals
  - **Hypothesis Testing**

    Wants to make a decision about the value of the parameter

    T-tests, F-tests
- Both based on premise of repeat experiments
- Will incorporate both in regression/ANOVA

Estimation

- Consider unknown pop parameter θ
- Select a random sample of indivs from the pop
- Will estimate using numerical summary $\widehat{\theta}$
- Summary known as a **point estimate** (single value)

  Examples: $\overline{X} \rightarrow \mu$ and $S^2 \rightarrow \sigma^2$

  Both estimates unbiased $\rightarrow$ mean of $\widehat{\theta}$ is θ
- Point estimate gives no indication of precision
- Use std error to construct **confidence interval**

  Common form $\widehat{\theta} \pm t_{\alpha/2} S_{\widehat{\theta}}$

  In long run, θ in 100(1-α)% of the intervals

  100(1-α)% confident single CI contains θ

  Larger confidence level $\rightarrow$ wider interval

Examples

- **Problem 13, Chpt 3** A random sample of $n = 32$ people attending a diet clinic (initial 3 weeks) resulted in a weight loss average of $\overline{x} = 30.0$ lbs and std deviation of $s = 11$ lbs. Interest is in estimating the average weight loss during these weeks at this clinic. The form of the confidence interval is $\overline{x} \pm t_{\alpha/2,n-1} s/\sqrt{n}$. For a 99% CI, $t = 2.75$ (using df=30). Thus, the CI is $30 \pm 2.75(11/\sqrt{32}) = (24.65, 35.35)$.
- A fishery is interested in the weight gain of trout in relation to two factors, water temperature and water movement. Specifically, they are interested in assessing whether the difference in weight gain between still and flowing water is the same in both warm and cold water. To study this, the following four treatments were randomly assigned equally to 20 one year-old fish. Diet and other conditions are held as constant as possible. The weight gain over a six month period was measured.

Trt	Description	Sample Mean (lbs)
1	Cold and Still	2.50
2	Cold and Flowing	3.00
3	Warm and Still	2.25
4	Warm and Flowing	2.30

The four sample variances (not shown) were pooled together to give $S_p = 0.27$ with 16 df. The estimated difference in cold water is $3.00 - 2.50 = 0.50$ and the estimated difference in warm water is $2.30 - 2.25 = 0.05$. Thus the difference between cold water and warm water is $0.50 - 0.05 = 0.45$. The form of the confidence interval is $\sum c_i \overline{x}_i \pm t_{\alpha/2,\text{df}} S_p \sqrt{\sum c_i^2/n_i}$. Here, df=16, $n_i = 5$, $S_p = .27$, $c_1 = -1, c_2 = 1, c_3 = 1$, and $c_4 = -1$. A 95% CI for this comparison of warm and cold water is $0.45 \pm 2.120(.27)\sqrt{4/5} = (-0.06, 0.96)$. While the observed difference is bigger in cold water, the 95% CI does contain zero.

Hypothesis Testing

- CI gives region of "likely" values for the parameter
- Often want to test if parameter in a specific region
- General procedure
  1. Look at data, check assumptions
  2. State the null (H_0) and alternative hypotheses (H_A)
  3. Specify significance level
  4. Specify test statistic and its sampling dist under H_0
  5. Form decision rule
  6. Compute statistic and draw conclusion
- Similar to CI, sampling dist based on repeat trials
- P-value quantifies how unusual $\hat{\theta}$ is under H_0

Two-sample T-Test

- $H_0 : \mu_1 = \mu_2$ (Null Hypothesis)
- $H_A : \mu_1 \begin{matrix} < \\ > \\ \neq \end{matrix} \mu_2$ (Alternative Hypothesis)
- Collect data - n_1 and n_2 observations

$$x_{11}, x_{12}, \ldots, x_{1n_1}$$
$$x_{21}, x_{22}, \ldots, x_{2n_2}$$

$$\overline{x}_1 = \frac{x_{11} + \ldots + x_{1n_1}}{n_1} \qquad \overline{x}_2 = \frac{x_{21} + \ldots + x_{2n_2}}{n_2}$$

- Is observed difference $\overline{x}_1 - \overline{x}_2$ "unusual" if $\mu_1 = \mu_2$?

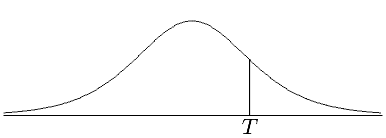
Use $T = (\overline{x}_1 - \overline{x}_2)/\mathrm{S_p}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where

$$\mathrm{S_p^2} = \frac{(n_1-1)\mathrm{S_1^2} + (n_2-1)\mathrm{S_2^2}}{n_1+n_2-2}$$

Assumptions

1. Independent observations
2. Equal variances
3. Normally distributed observations

- Assuming H_0: $\mu_1 = \mu_2$ and these three assumptions. the distribution of T is t-distributed with $n_1 + n_2 - 2$ degrees of freedom
- "Unusual" then quantified by the probability that a randomly drawn t is more extreme then T (tail region of distribution)
- Reject null hypothesis if this probability is "small". "Small" based on choice of significance level α. One-sided example shown below.



Example

In a study of lettuce growth, **ten seedlings were randomly allocated** to be grown in either **a standard nutrient solution** or in **a solution containing extra nitrogen**. After 22 days, the plants were harvested and weighed. The table below summarizes the results. Can we conclude that extra nitrogen enhances growth?

	Leaf Dry Weight (gm)		
Nutrient Solution	n	Mean	SD
Standard	5	3.62	0.54
Extra	5	4.17	0.67

Solution: Do not have the data to inspect. Assume assumptions satisfied. The alternative hypothesis is $H_A : \mu_E > \mu_S$. The pooled variance $\mathrm{S_p^2} = (4(.54)^2 + 4(.67)^2)/8 = 0.37$. Our test statistic is then $T = (4.17 - 3.62)/\sqrt{2(.37)/5} = 1.43$. With 8 degrees of freedom, the P-value is between .05 and .10. From Table A-2, $t_{.10} = 1.397$ and $t_{.05} = 1.860$. If α were greater than .10, we would reject the null and conclude that extra nitrogen enhances growth. If α were less than .05, we would **not reject the null** and conclude there is not sufficient evidence to state that the extra nitrogen enhances growth.

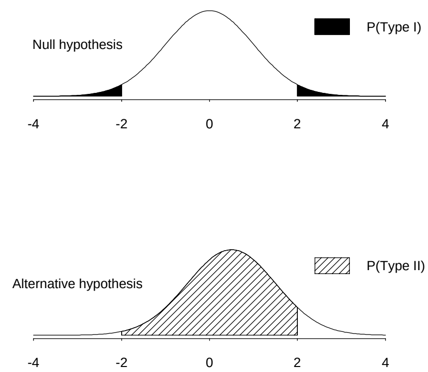
Type I and Type II errors

- In hypothesis testing, two types of errors

		TEST RESULT DNR	TEST RESULT R
REALITY	H_0	♡	I
	H_A	II	♡

- **Type I error**: α = P(reject $H_0 | H_0$ true)
- **Type II error**: β = P(do not reject $H_0 | H_0$ false)
- Power of test (for specific H_A) is $1 - \beta$
- Significance level is α (this defines "unusual")

Choice of Sample Size/Computing Power



- **Goal of test:** Detect diff of size Δ with high prob

$$|\mu_1 - \mu_2| = \Delta$$

- Choice of Δ subjective (practical significance)
- Probability to detect difference is power
- Power depends on α, Δ, σ, and n

Power/Sample Size Calculations

- Assume $\neq$ alternative w/ σ known and $n_1 = n_2 = n$

$H_0 : \overline{X}_1 - \overline{X}_2 \frown N(0, \sigma\sqrt{2/n})$

$H_A : \overline{X}_1 - \overline{X}_2 \frown N(\Delta, \sigma\sqrt{2/n})$

- Reject if (use H_0 dist)

$$\overline{X}_1 - \overline{X}_2 > z_{\alpha/2}\sigma\sqrt{2/n}$$
or
$$\overline{X}_1 - \overline{X}_2 < -z_{\alpha/2}\sigma\sqrt{2/n}$$

- Power: P(Reject when H_A true) (use H_A dist)

  After standardization, simplifies to

  $P(Z > z_{\alpha/2} - \Delta/\sqrt{2\sigma^2/n}) + P(Z < -z_{\alpha/2} - \Delta/\sqrt{2\sigma^2/n})$

- This gives power for particular n
- Can also specify power and solve for n (pg 29)

Power Calculations (σ unknown)

$H_0 : |\mu_1 - \mu_2| = 0$

$H_A : |\mu_1 - \mu_2| = \Delta$

Reject if:

$$\overline{X}_1 - \overline{X}_2 > t_{2(n-1),1-\alpha/2}\sqrt{2S_p^2/n}$$
or
$$\overline{X}_1 - \overline{X}_2 < t_{2(n-1),\alpha/2}\sqrt{2S_p^2/n}$$

Power: P(reject | H_A)

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{2S_p^2/n}} \sim t_{2(n-1)}(\Delta/\sqrt{2\sigma^2/n})$$

Noncentral parameter $\Delta/\sqrt{2\sigma^2/n}$

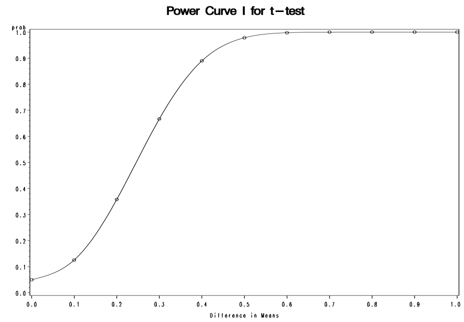
Compute probability of rejection given noncentral t

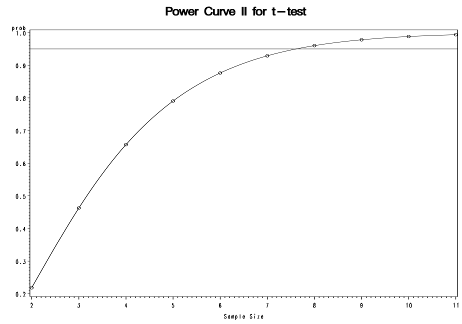
Using SAS

tpower.sas

```
options nocenter;
goptions htext=1.0 htitle=1.5 ftext=swiss ftitle=swissb cback=white
        hsize=5in vsize=4in vpos=35 hpos=35;
/* Compute power curve for the following parameters */
data new;
 input n alpha sigma;
 cards;
   9 .05 .25
 ;

data new1; set new;
 do delta = 0 to 1 by .10;
    df = 2*(n-1); nc = delta/(sigma*sqrt(2/n));
    rlow  = tinv(alpha/2,df); rhigh = tinv(1-alpha/2,df);
    p=1-probt(rhigh,df,nc)+probt(rlow,df,nc); output;
 end;

symbol1 v=circle i=sm5; title1 'Power Curve I for t-test';
axis1 label=('prob'); axis2 label=('Difference in Means');
proc gplot;  plot p*delta / haxis=axis2 vaxis=axis1; run;
/* Find an appropriate sample size for 95% power at delta =.5 */
data new2; set new;
 d1=.5;
 do n1=2 to 11 by 1;
    df = 2*(n1-1); nc = d1/(sigma*sqrt(2/n1));
    rlow  = tinv(alpha/2,df); rhigh = tinv(1-alpha/2,df);
    p=1-probt(rhigh,df,nc)+probt(rlow,df,nc); output;
 end;

symbol1 v=circle i=sm5; title1 'Power Curve II for t-test';
axis1 label=('prob'); axis2 label=('Sample Size');
proc gplot;  plot p*n1 / haxis=axis2 vaxis=axis1 vref=0.95; run;
```


Ouput





Paired Comparison

- Can often improve precision by pairing
- Removes explainable variation from the analysis
- Like material in each population

  Twins for drug/health studies

  Same specimen given both trts

  Similar plots in a field
- Look at difference between **each** pair
- Changing $2n$ observations into n indep obs

$$d_i = y_{1i} - y_{2i}$$
$$S_d^2 = \frac{1}{n-1}\sum(d_i - \overline{d})^2$$

$$T = \overline{d}/(S_d/\sqrt{n})$$
$$T \sim t_{n-1}$$

Example

Paired T-test/Randomization Paired Test

In a study of egg cell maturation, the eggs from each of four female frogs were divided into two batches and one batch was exposed to progesterone. After two minutes, the cAMP content was measured. It is believed that cAMP is a substance that can mediate cellular response to hormones.

FROG	cAMP Content		
	Control	Progesterone	Diff
1	6	4	2
2	4	5	-1
3	5	2	3
4	4	2	2

- **t-test:** $d = \{2, -1, 3, 2\} \to \overline{d} = 1.5$ and $s_{\overline{d}} = .866$. The test statistic is 1.732. Using Table II and 3 degrees of freedom, the P-value is between .05 and .10 (one-sided), .10 and .20 (two-sided). The actual two-sided P-value is close to 0.18.
- **randomization:** The result of each pair does not depend on the allocation of treatments. Thus there are $2^4 = 16$ possible outcomes. The observed outcome is 2-1+3+2=6. The combinations 2+1+3+2 and -2-1-3-2 give the values of 8 below.

$\lvert\sum d\rvert$	# of occurrences
8	2
6	2
4	4
2	6
0	2

From the table, there are four of sixteen outcomes as or more "unlikely" simply due to chance. Thus the P-value is 0.25.

CIs vs Hypothesis Tests

- Consider two-sided hypothesis test w/ level α
  - Reject if $|\overline{X}_1 - \overline{X}_2| > t_{\alpha/2} S_p \sqrt{1/n_1 + 1/n_2}$
- Consider $100(1-\alpha)\%$ CI
  - Half-width of CI is $t_{\alpha/2} S_p \sqrt{1/n_1 + 1/n_2}$
  - 0 not in interval if $|\overline{X}_1 - \overline{X}_2| > t_{\alpha/2} S_p \sqrt{1/n_1 + 1/n_2}$
- Will reject H_0 if 0 not in confidence interval
- Can immediately test any $H_0 : \Delta = \Delta_0$ at level α

Multiple Testing

- Often perform more than one test on data set
- If each test uses significance level α

  P(at least one Type I error) $> \alpha$
- Various way to control overall Type I error rate
- Bonferroni's Correction

  Recall P($A \cup B$) = P(A) + P(B) - P($A \cap B$)

  P(at least one type I error in k tests) $\leq k\alpha$

  For each test use $\alpha' = \alpha/k$

  Then, P(at least one type I error in k tests) $\leq \alpha$

  Designed for planned comparisons only

  Extremely conservative if k is large