STOR 455
STATISTICAL METHODS I

Jan Hannig

Some SAS

```
data blah;
infile 'C:\Documents and Settings\Administrator\My Documents\MyDropbox\SAS\stuff\AGE18.DAT';
input x1-x8;
run;

proc print data =blah;
run;

proc means data = blah;
var x2-x3;
run;
```

Inferential Procedures

- Once the sample is selected, inference about the sample is performed
- Types of inference
  - Point Estimation
  - Confidence Intervals
  - Hypothesis Testing

Confidence Interval

- *1-α*: confidence level,
  - how confident we are that the confidence interval will cover the true population mean.
- Want to find a level C=*1-α* confidence interval for θ
- Such that P($L \le \theta \le U$)=1-α.
  - The meaning of this is "repeated sampling" probability [see (1.6.5) in the book]
  - In many applications we have

$$\hat{\theta} - \text{table value} \cdot SE(\hat{\theta}) \le \theta \le \hat{\theta} + \text{table value} \cdot SE(\hat{\theta})$$

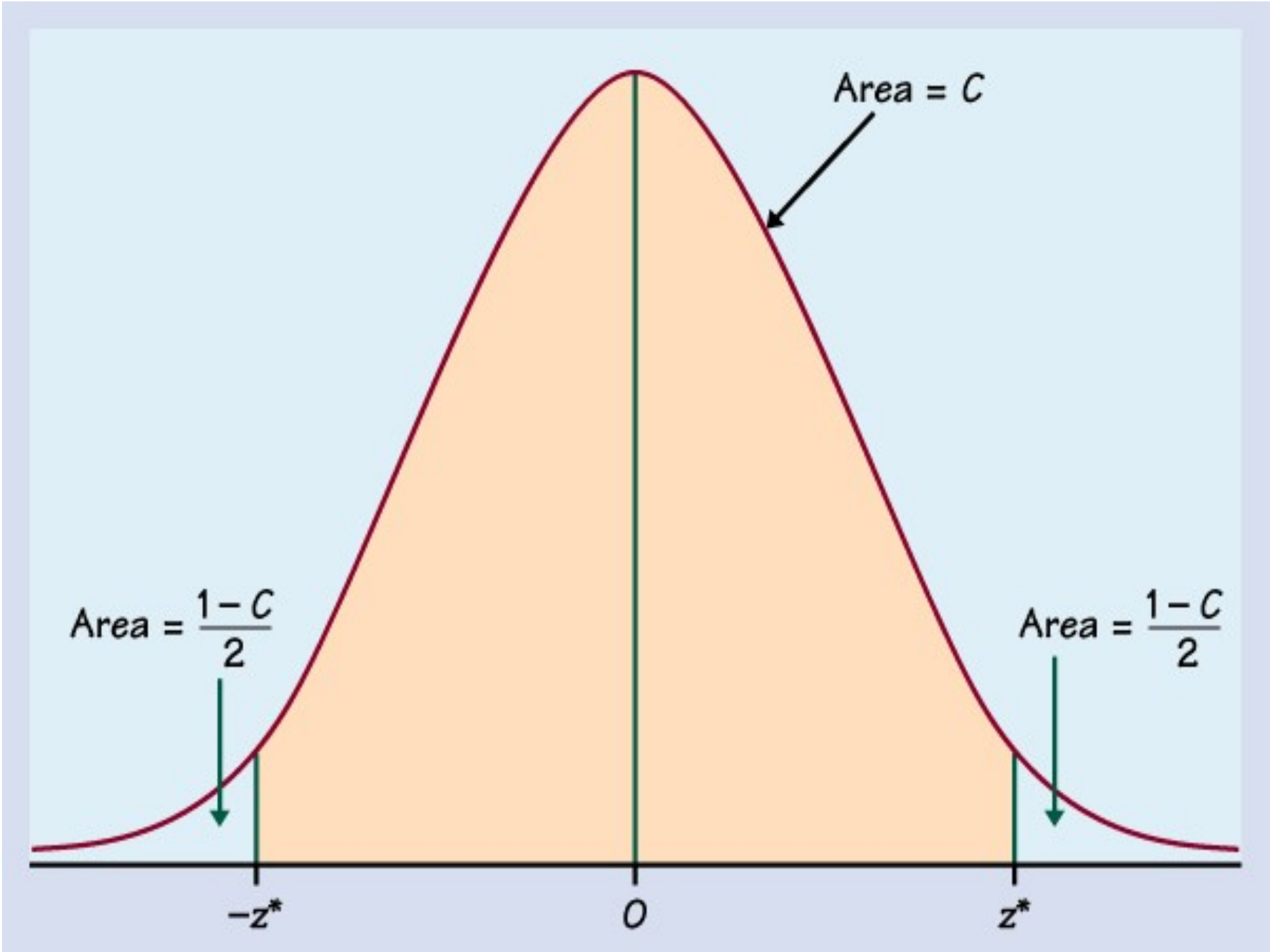
  - *SE* is the *standard error (estimate of the sd)* of the point estimator.

Confidence Interval for μ

- To estimate μ, a sample of size n is drawn from the population, and its mean $\bar{x}$ is calculated.
- We know that $\bar{x}$ is (approximately) normally distributed, and

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1).$$

- σ is known


Area = C
Area = $\frac{1-C}{2}$
Area = $\frac{1-C}{2}$
$-z^*$
0
z^*

- Then,

$$P(-z_* \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_*) = 1-\alpha$$

$$P(\mu - z_* \frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + z_* \frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

- This leads to

$$P(\bar{x} - z_* \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_* \frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

- Thus, the level *C=1-α* confidence interval for μ is

$$[\bar{x} - z_* \frac{\sigma}{\sqrt{n}}, \bar{x} + z_* \frac{\sigma}{\sqrt{n}}]$$

Inference for a normal mean μ with unknown σ

- For CI and hypothesis testing about a normal mean μ, when σ is *not known*, the sample standard deviation *s* is used to estimate σ.

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \qquad t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

- *t* has a Student *t* distribution with *n* - 1 degrees of freedom (df).
- s is the sample standard deviation $s = \sqrt{\frac{SSX}{n-1}}$
- $s/\sqrt{n}$ is the standard error

Confidence Interval for μ

- A 100(1 - α)% confidence interval for μ is

$$[\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}]$$

- or, more compactly,

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}.$$

Investment Return

- An investor is estimating the return on investment in companies that won quality awards last year.
- A random sample of 50 companies is selected, and the return on investment is calculated had he invested in them. The data is summarized as follows:

$$\bar{x} = 14.75, \quad s = 8.18.$$

- Construct a 95% CI for the mean return.

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 14.75 \pm 2.009\frac{8.18}{\sqrt{50}} = \left[12.43, 17.07\right]$$

SAS Example

- `proc means data = xxx alpha=0.05 clm mean std;`
- `var xxx xxx;`
- `run;`

The SAS System 16:36 Tuesday, August 31, 2010 2

The MEANS Procedure

Analysis Variable : VAR3

Lower 95% CL for Mean	Upper 95% CL for Mean	Mean	Std Dev
1.3743701	4.8256299	3.1000000	2.7159462

Prediction Intervals

- Sometimes we do not want to estimate a parameter. Instead we want to estimate the future data value.
- The confidence interval is

$$\hat{Y}_f - \text{table value} \cdot SE(Y_f) \leq Y_f \leq \hat{Y}_f + \text{table value} \cdot SE(Y_f)$$

- If Y are N(μ,σ) then "table value" is t and

$$\hat{Y}_f = \bar{y}, \quad SE(Y_f) = s\sqrt{1 + \tfrac{1}{n}}$$

Investment Return

- An investor is estimating the return on investment in companies that won quality awards last year.
- A random sample of 50 companies is selected, and the return on investment is calculated had he invested in them. The data is summarized as follows:

$$\bar{x} = 14.75, \quad s = 8.18.$$

- Construct a 95% PI for another company that won an award.

$$\bar{x} \pm t_{\alpha/2,n-1} s\sqrt{1+\frac{1}{n}} = 14.75 \pm 2.0098.8.18\sqrt{1+\frac{1}{50}} = \left[-1.85, 31.35\right]$$

Summary of CI

Notation: $\bar{y} = \frac{1}{n}\sum y_i;\ SSY = \sum (y_i - \bar{y})^2$	
Inference	**Formulas and Procedures**
Point estimate of μ_Y, σ_Y	$\hat{\mu}_Y = \bar{y}$ $\hat{\sigma}_Y = \sqrt{SSY/(n-1)}$
Two-sided $1-\alpha$ confidence intervals for μ_Y	$\hat{\mu}_Y - t_{1-\alpha/2:n-1} SE(\hat{\mu}_Y) \le \mu_Y \le \hat{\mu}_Y + t_{1-\alpha/2:n-1} SE(\hat{\mu}_Y)$ where $SE(\hat{\mu}_Y) = \frac{\hat{\sigma}_Y}{\sqrt{n}}$
Two-sided $1-\alpha$ confidence intervals for σ_Y	$\sqrt{\frac{SSY}{\chi^2_{1-\alpha/2:n-1}}} \le \sigma_Y \le \sqrt{\frac{SSY}{\chi^2_{\alpha/2:n-1}}}$
Two-sided $1-\alpha$ confidence intervals for Y_0	$\hat{\mu}_Y - t_{1-\alpha/2:n-1}\hat{\sigma}_Y\sqrt{1+\frac{1}{n}} \le Y_0 \le \hat{\mu}_Y + t_{1-\alpha/2:n-1}\hat{\sigma}_Y\sqrt{1+\frac{1}{n}}$

Concepts of Hypothesis Testing

- Two hypotheses:
  - H_0: the null hypothesis
    - The statement of "no effect" or "no difference".
    - The statement we try to find evidence against.
  - H_a: the alternative hypothesis
    - The statement we hope or suspect is true.

Z test

- Problem of interest:
  - Population mean μ of a normal distribution
  - known σ
- Null hypothesis: H_0: $\mu = \mu_0$
- Test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}.$$

H_a	p-value
$\mu \neq \mu_0$	$2P(Z \geq \lvert z \rvert)$
$\mu > \mu_0$	$P(Z \geq z)$
$\mu < \mu_0$	$P(Z \leq z)$

t Test

- Problem of interest:
  - Population mean μ of a normal distribution
  - Unknown σ - estimated by s
- Null hypothesis: H_0: $\mu = \mu_0$
- Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.$$

H_a	p-value
$\mu \neq \mu_0$	$2P(T \geq \lvert t \rvert)$
$\mu > \mu_0$	$P(T \geq t)$
$\mu < \mu_0$	$P(T \leq t)$

P-value

- The repeated sampling probability of observing a test statistic as extreme or more extreme than the actually observed value, under the cond. that H_0 is true.
  - "extreme" means "far from what we expect from H_0".
  - The smaller a *P-value*, the less evidence for H_0, or the more evidence for H_a
  - *P-value* provides information about the amount of statistical evidence that supports the null hypothesis.

Describing *P-value*

- If the *P-value* is less than 1%, there is ***overwhelming evidence*** that supports the alternative hypothesis.
- If the *P-value* is between 1% and 5%, there is ***strong evidence*** that supports the alternative hypothesis.
- If the *P-value* is between 5% and 10% there is ***weak evidence*** that supports the alternative hypothesis.
- If the *P-value* exceeds 10%, there is ***no evidence*** that supports of the alternative hypothesis.

Significance Level α and Statistical Significance

- We need to make a conclusion after carrying out the hypothesis test. *What do we conclude?*
- We compare the *P-value* with a fixed value that we regard as decisive.
- This amounts to deciding in advance how much evidence against H_0 we require in order to reject H_0.
- The decisive value is called the *significance level* of the test. It is denoted by α and the corresponding test is called a *level α test*.

Statistical Significance: If the *P-value* ≤ α, we say that the data are statistically significant at level α.

α and *P-value*

- *P-value* and significance level α:
  - Reject H_0 if $p \leq \alpha$
  - Do not reject H_0 if $p > \alpha$.
- When is the evidence against H_0 stronger?
  - *Large P-value or small P-value?*
  - The smaller the *P-value*, the stronger the evidence against H_0 and in favor of the alternative H_a.
- When is it easier to reject H_0?
  - *Large α or small α ?*
  - We need a lot more evidence to reject H_0 for small α than for large α.

Four steps of hypotheses testing

- Define the hypotheses to test, and the significance level α.
- Calculate the value of the test statistic.
- Find the *P-value* based on the observed data.
- State the conclusion.
  - Reject the null hypothesis if the *P-value* <=α ; if it>α, the data do not provide sufficient evidence to reject the null.

Productivity of Newly-hired

- To determine number of workers required to meet demand, the productivity of new trainees is studied.
- It is believed trainees can process more than 450 packages per hour within one week of hiring.
- A sample of productivity of 50 trainees is observed and summarized as:

$$\bar{x} = 460.38, \quad s = 38.83.$$

- Can we conclude that this belief is correct based on the sample?

Productivity of Newly-hired

- The problem objective is to describe the population of the number of packages processed in one hour.
- The hypotheses are

  H_0: $\mu = 450$ vs. H_a: $\mu > 450$
- The t statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{460.38 - 450}{38.83/\sqrt{50}} = 1.89$$

- d.f. = n - 1 = 49, p-value = 0.0323.

- There is sufficient evidence to infer that the mean productivity of trainees one week after being hired is greater than 450 packages at .05 significance level.

CI & 2-Sided Tests

- A level α 2-sided test
  - Accepts H_0: $\mu = \mu_0$ exactly when the value μ_0 falls inside a level $1 - \alpha$ confidence interval for μ.
  - rejects H_0 when the value μ_0 falls outside the CI.
- CI can be used to test hypotheses.
  - Calculate the $1 - \alpha$ level confidence interval, then
    - if μ_0 falls within the interval, accept the null hypothesis,
    - Otherwise, reject the null hypothesis.
- We *can* obtain a result of a *test from CI*.
- We *cannot* obtain *CI from* a result of one *test*

Reaction Time to Traffic Signs

- To compare mean reaction times to two types of traffic signs,
  - prohibitive (I) (e.g. No Left Turn)
  - permissive (II) (e.g. Left Turn Only)
- 15 subjects were chosen and each subject was presented with 40 traffic signs, 20 prohibitive and 20 permissive, in random order. The mean reaction times of the 15 subjects are as follows:

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
I	7.6	10.2	9.5	1.3	3.0	6.3	5.3	6.2	2.2	4.8	11.3	12.1	6.9	7.6	8.4
II	7.3	9.1	8.4	1.5	2.7	5.8	4.9	5.3	2.0	4.2	11.0	11.0	6.1	6.7	7.5
d	0.3	1.1	1.1	-0.2	0.3	0.5	0.4	0.9	0.2	0.6	0.3	1.1	0.8	0.9	0.9

Remark

- There is significant variability between subjects. Some people react more quickly to any type of sign than other people;

- The correlation within subjects is large. If a person reacts quickly to one type of signs, he/she is more likely to react quickly to another type of signs as well.

Compare Two populations

- *Matched pairs*: Parameter: μ_1 - μ_2
- *Compute differences* and use usual t test on the differences.
- Condition: *the differences are normally distributed*.
  - Test statistic:

$$t = \frac{\bar{d} - \Delta_0}{s_d / \sqrt{n}}, \quad \text{d.f.} = n - 1.$$

  - Confidence interval:

$$\bar{d} \pm t_{\alpha/2, n-1} \cdot s_d / \sqrt{n}.$$

SAS Example

```
data blah1;
   set ABSORPT;
   d=Column_2-Column_1;
run;
proc means data=blah1 mean std clm prt;
var d;
run;
```

The SAS System 16:36 Tuesday, August 31, 2010 7

The MEANS Procedure

Analysis Variable : d

Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean	Pr > |t|
0.1500000	4.0495791	-2.4229798	2.7229798	0.9002

Simultaneous Tests

- A company has developed four new additives that can be added into cement. An experiment has been carried out to determine if there are any differences between the average strength of cement blocks made with each additive.
- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ H_A: there is a difference
- We are testing more than one difference at a time!
- H_0: $\theta_1 = \mu_2 - \mu_1 = 0$, $\theta_2 = \mu_3 - \mu_1 = 0$, $\theta_3 = \mu_4 - \mu_1 = 0$

Bonferroni Adjustment

- Divide your α between the test.
- In our example we have 3 tests – use $\alpha/3$ as a significance level.

READ THE BOOK. (Conversation 1.6 in particular)