Introduction to statistics
Lecture 22
Tuesday, Nov. 22, 2005

- I will archive all project forums tomorrow morning for grading!
- Late project? Five-point penalty.

Sampling distribution model of sample means
- By CLT, if population standard deviation is known.
  \[ \bar{X} - \mu \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \]
- What if \( \sigma \) (population std. dev) is unknown:
  \[ \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim t \text{ distribution with } (n-1) \text{ d.f.} \]

The sample estimation of the standard deviation

What is a t distribution
- An SRS of \( n \) observations from a normal distribution \( N(\mu, \sigma) \)
- The sampling distribution of \( \frac{\bar{X} - \mu}{S/\sqrt{n}} \)
  is called a t distribution
  with (n-1) degrees of freedom.
- This is not an equivalency but a definition.
- (n-1) is the parameter of the t distributions.
- t distribution can be viewed as a standard normal distribution that got pressed down.

We say t distributions have fatter (heavier) tails!
It means it is more likely to observe values that are far from the mean under the t distributions than under a standard normal distribution.

What about two sample means
- If \( \sigma_1, \sigma_2 \) are known
  \[ \bar{X}_1 - N(\mu_1, \frac{\sigma_1^2}{n_1}) \]
  \[ \bar{X}_2 - N(\mu_2, \frac{\sigma_2^2}{n_2}) \]
  \[ (\bar{X}_1 - \bar{X}_2) - N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \]
- Z procedures for inference are based on
  \[ \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \]

What if the population standard deviations are unknown
- \( t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \)
  can be approximated by a t-distribution with a carefully calculated degrees of freedom.
  When doing things by hand, to simplify, we then use a conservative distribution: the t distribution with [the smaller of (n1-1) and (n2-1)] degrees of freedom.
  See page 480 on “An Easier Rule?” in the side note.
Inference on population means

- Basis of inference: an estimator, sample mean, $x$-bar.
- Understanding of its sampling variability: the sampling distribution model of $x$-bar
- Level of uncertainty you are willing to compromise:
  - choosing imperfect level of confidence,
  - non-zero significance level which allows some type I error rejecting the null hypothesis.

Critical values of t distribution

- For the same confidence level, the critical value from $N(0,1), Z^*$, is always smaller than the critical value from a t-distribution.
- The lower the degrees of freedom is, the further $t^*$ from the corresponding $z^*$ (at a same confidence or significance level).
- But the critical values are similarly defined under different distribution models.

CI, standard deviation known

- The confidence interval is defined as
  Estimate $\pm$ margin of error
- Margin of error
  - Critical value by Confidence level: ‘a standard width’ decided by the levels of confidence
  - Sampling variability: standard deviation of the sampling distribution

  $$
  \bar{X} \pm z^* \frac{\sigma}{\sqrt{n}}
  $$

Hypothesis test, std dev known

- Null hypothesis: statement about $\mu$
  $$
  H_0 : \mu = \mu_0
  $$
- Two sided alternative or one-sided alternative
  $$
  H_a : \mu \neq \mu_0 \quad \text{or} \quad H_a : \mu > \mu_0 \quad \text{or} \quad H_a : \mu < \mu_0
  $$
- Observed value: sample mean; Sampling distribution of sample mean under the null hypothesis:
  $$
  N(\mu_0, \frac{\sigma}{\sqrt{n}})
  $$
- Test statistic
  $$
  z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}
  $$

One-sample t CI

- Std. Dev. Unknown
- Confidence interval:
  estimate $\pm$ margin of error
- Sampling distribution for
  $$
  t = \frac{\bar{X} - \mu}{s/\sqrt{n}}
  $$
  is $t$ distribution with $n-1$ degrees of freedom.
- Using critical values $t^*$ from $t$ distributions
  $$
  \bar{X} \pm t^* \frac{s}{\sqrt{n}}
  $$

One-sample t test

- Std. Dev. Unknown
- Null hypothesis: statement about $\mu$
  $$
  H_0 : \mu = \mu_0
  $$
- Two sided alternative and one-sided alternative
  $$
  H_a : \mu \neq \mu_0 \quad \text{or} \quad H_a : \mu > \mu_0 \quad \text{or} \quad H_a : \mu < \mu_0
  $$
- Observed value: sample mean; Sampling distribution of sample mean under the null hypothesis standardized using estimated standard deviation:
  $$
  t(n-1)
  $$
- Test statistic
  $$
  t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}
  $$
  and compare it with $t(n-1)$ to derive the p-value.
Find the t distribution first and then find the p-value

STATA for one sample mean
- See handout.

Comparing two means
- Data
  - Sample 1: $n_1$ obs. from population 1 ($\mu_1, \sigma_1$)
  - Sample 2: $n_2$ obs. from population 2 ($\mu_2, \sigma_2$)
- One then can calculate that

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\bar{X}_1$</td>
<td>$S_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\bar{X}_2$</td>
<td>$S_2$</td>
</tr>
</tbody>
</table>

$\bar{X}_1 - \bar{X}_2$ naturally capture the differences between the means.

t procedures
- Confidence interval
  $$(\bar{X}_1 - \bar{X}_2) \pm t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$
- Significance test
  $H_0: \mu_1 = \mu_2$
  $H_a: \mu_1 \neq \mu_2$; $H_a: \mu_1 > \mu_2$; $H_a: \mu_1 < \mu_2$
  test statistic: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
- t distribution used: t distribution with the [smaller of $(n_1-1)$ and $(n_2-1)$] degrees of freedom.

STATA for statistical inference
- See handout.

Try a Test

$n_1 = 8, \bar{X}_1 = 78.5, S_1 = 10$

$n_2 = 8, \bar{X}_2 = 83.5, S_2 = 10$

$H_0 : \mu_1 = \mu_2$?
The actual data

<table>
<thead>
<tr>
<th>Students</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>77</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>79</td>
</tr>
<tr>
<td>5</td>
<td>77</td>
<td>82</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>91</td>
<td>96</td>
</tr>
<tr>
<td>8</td>
<td>93</td>
<td>98</td>
</tr>
</tbody>
</table>

Reading

- Chapters 23-25
- Enjoy your Turkey with some statistics and come back for more!
- Happy holiday!