ECO220Y

Estimation: Confidence Interval Estimator for Population Mean

Readings: Chapter 13 (sections 13.1-13.4)

Fall 2011

Lecture 12

Wendy's vs. McDonald's

Between Wendy's and McDonald's, which fast-food drive-through window is faster? To answer the question, a random sample of service times for each restaurant was measured. If in fact there is no difference in the average service times between the two, what is the chance that McDonald's sample average service time would be higher? [Before we even collected the data]

The data:

	Wendy's	McDonalds
n	213	202
$\bar{X}$	150	154
σ_x	21	24

Wendy's vs. McDonald's

- Statistically, we would like to find the chance that the difference between two sample means is greater than 0.
- $P(\bar{X}_M - \bar{X}_W > 0) = ?$
- We need to know the distribution of $\bar{X}_M - \bar{X}_W$ to standardize and find the value from the standard normal table.
- Last time, we showed that $\bar{X}_W - \bar{X}_M \sim N(\mu_M - \mu_W, 4.92)$ (Review Lecture 11)
- $P(\bar{X}_M - \bar{X}_W > 0 | \mu_W - \mu_M, \sigma_{\bar{x}_M - \bar{x}_W} = \sqrt{4.92}) = ?$
- Moreover, under assumption that there is no difference in service times between two fast-food chains, $\mu_M - \mu_W = 0$
- $\bar{X}_W - \bar{X}_M \sim N(0, 4.92)$

Wendy's vs. McDonald's

- $P(\bar{X}_M - \bar{X}_W > 0|0, \sqrt{4.92}) = P(Z > \frac{0-0}{\sqrt{4.92}}) = ?$
- What is the chance that sampling error explains the difference as large as the one we observe in the data?
- $P(\bar{X}_M - \bar{X}_W > 4|0, \sqrt{4.92}) = P(Z > \frac{4-0}{\sqrt{4.92}}) = P(Z > 1.80) = 0.0359$

Back to Reality

- In practice, we do not know population standard deviation, σ
- In order to make inference, we need to estimate this parameter together with μ.
- Obvious possibility: use sample standard deviation s to estimate σ.
- The need to estimate σ creates another source of uncertainty.
- Added uncertainty implies that we have to make confidence interval wider to allow for a little more margin for error.
- How to find the exact widening that is needed?

Student's t distribution

1876-1937

In 1908, W.S. Gosset figured out the exact widening that is needed. He was a statistician for the Irish brewery, Guinness, which encouraged statistical research but not publication. Gosset persuaded Guinness to allow his work to be published under the pseudonym "Student". His calculations became known as **Student's t distribution**.

σ is known

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

σ is unknown

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Standardize:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

Linear transformation of

one random variable

$$Z \sim N(0, 1)$$

$$Y = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

Linear transformation of

two random variables

$$Y \sim ?$$

t Distribution

- $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ is called t-statistic.
- t-statistic is a random variable.
- t-statistic is a function of two random variables. (??)
- The distribution of t-statistic depends on sample size n, because the reliability of the standard deviation estimate is greater when the sample size is larger.
- For an infinite sample $(n \rightarrow \infty)$, s should equal σ and the distribution of t and Z should coincide.

Student t Density Function

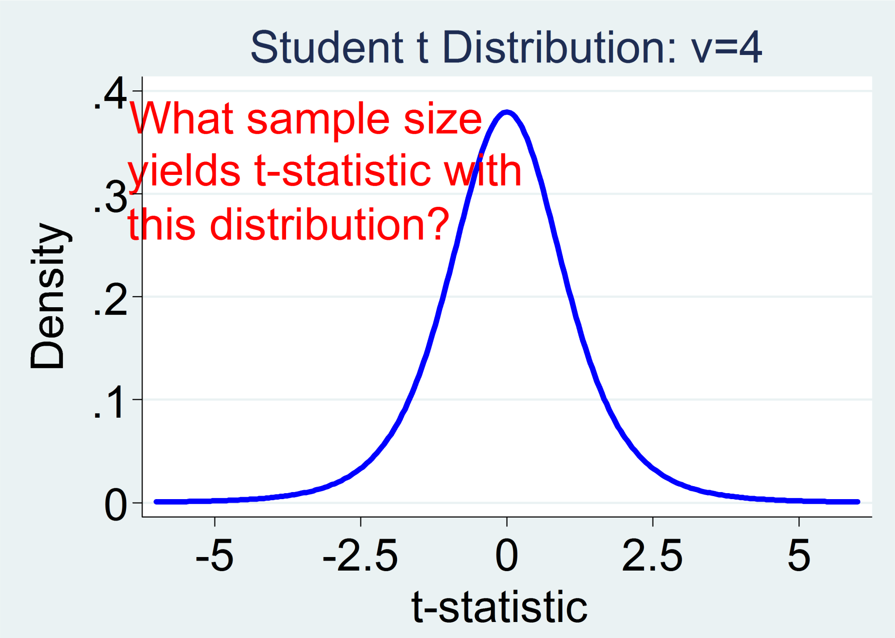
$$f(x) = \frac{\left[\frac{\nu-1}{2}\right]!}{\sqrt{\nu\pi}\left[\frac{\nu-2}{2}\right]!}\left[1 + \frac{x^2}{\nu}\right]^{\frac{-(\nu+1)}{2}}$$

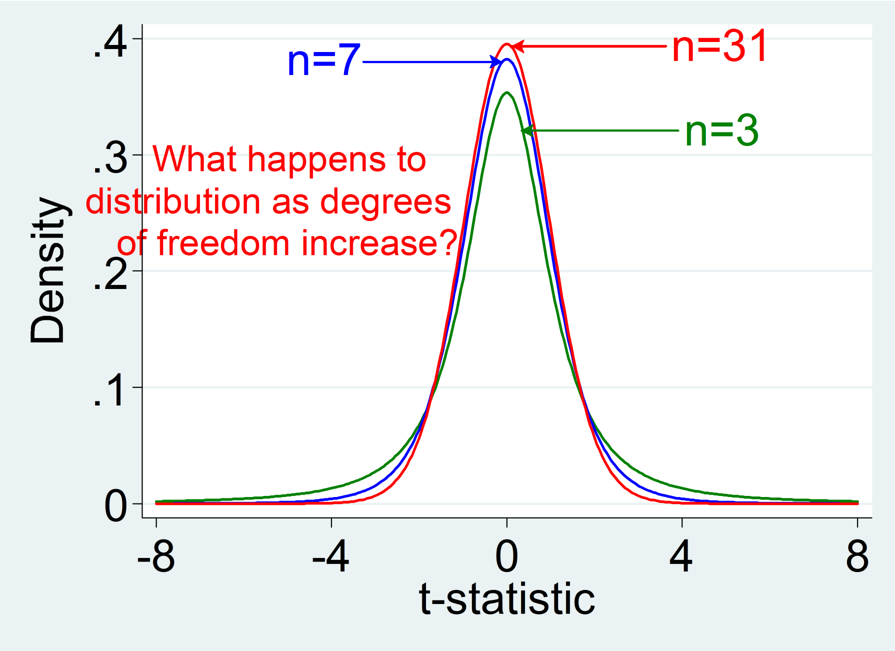
ν = 1, 2, 3, . . .

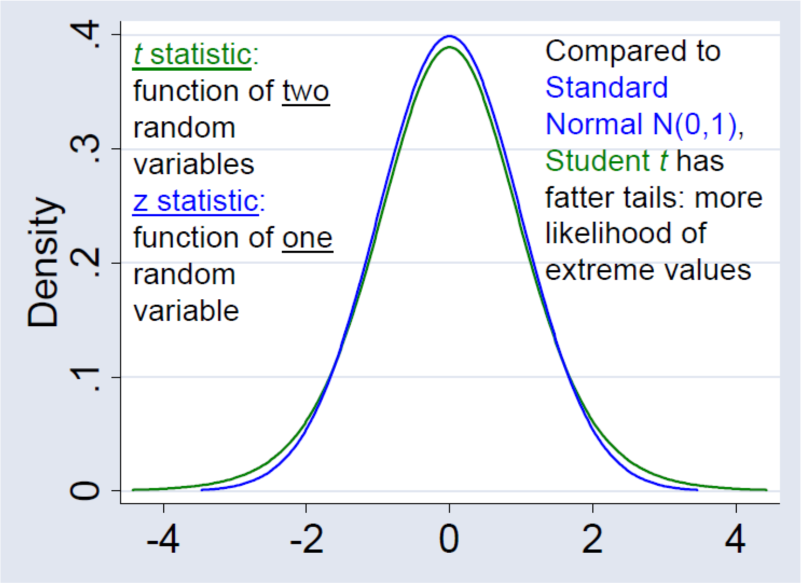
ν (read "nu") is called "degrees of freedom"

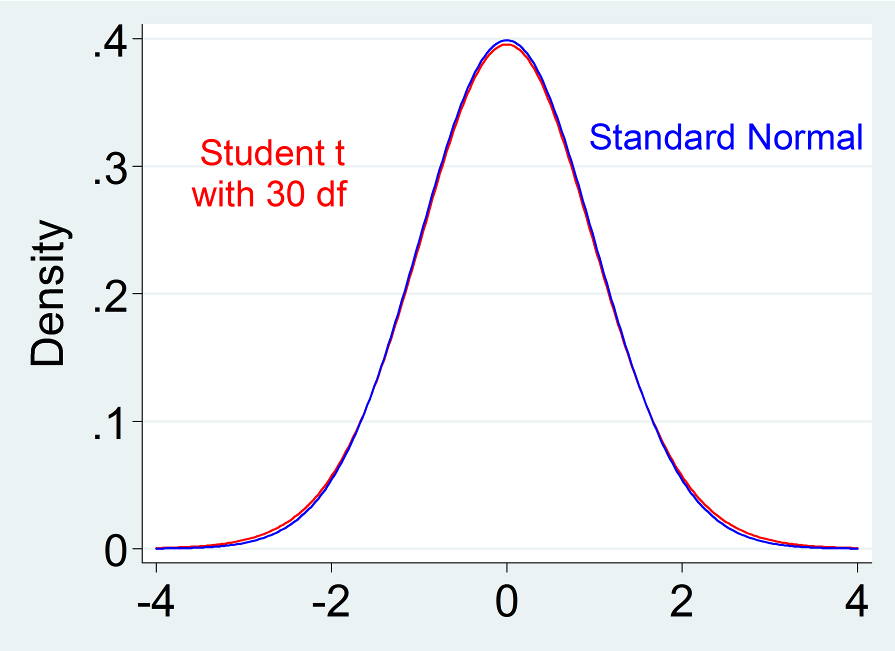
ν = n - 1

ν is the only parameter of Student t distribution


Student t Distribution: v=4
What sample size yields t-statistic with this distribution?
Density
.4
.3
.2
.1
0
-5
-2.5
0
2.5
5
t-statistic


n=7
n=31
n=3
What happens to
distribution as degrees
of freedom increase?
.4
.3
.2
.1
0
Density
-8
-4
0
4
8
t-statistic

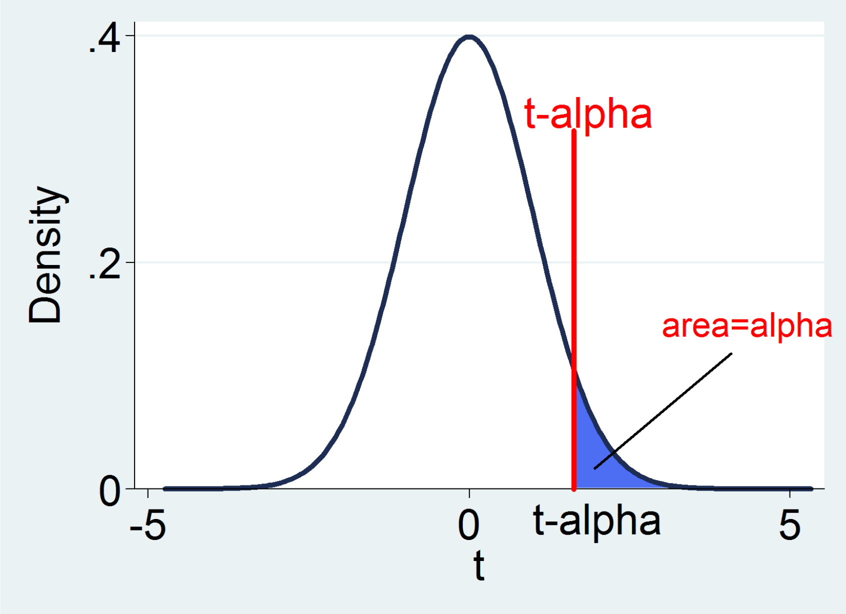

Density
0
.1
.2
.3
.4
-4
-2
0
2
4
t statistic:
function of two
random
variables
z statistic:
function of one
random
variable
Compared to
Standard
Normal N(0,1),
Student t has
fatter tails: more
likelihood of
extreme values


Density
.4
.3
.2
.1
0
-4
-2
0
2
4
Student t
with 30 df
Standard Normal

Finding Student t Probabilities

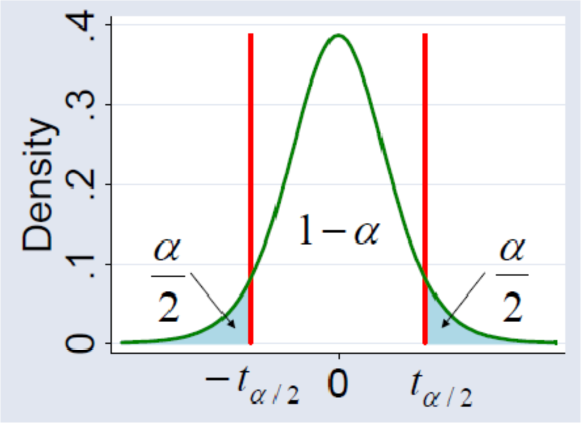
- Density function of t distribution is complex.
- Use tables to find cumulative probabilities.
- t-table distributed in class last week
  - Table reports P($T > t_A | \nu = n - 1$)=A
  - A=0.10, 0.05, 0.025, 0.01, 0.005
- When can we use normal probabilities?

$$P(T > t_\alpha | \nu = n - 1) = \alpha$$



Confidence Interval

$$P(-t_{\alpha/2} < t < t_{\alpha/2}) = 1 - \alpha$$



CI Estimator of μ when σ is unknown

- Confidence interval estimator of μ is:

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

- Confidence level: 1-α
- Lower confidence limit (LCL): $\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}$
- Upper confidence limit (UCL): $\bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$

Finding Student t Probabilities

$\nu = n - 1$	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657
10	1.372	1.812	2.228	2.764	3.169
30	1.310	1.697	2.042	2.457	2.750
200	1.286	1.653	1.972	2.345	2.601
∞	1.282	1.645	1.960	2.326	2.576

$P(T > 6.314 | \nu = 1) = 0.05$

$P(T > 2.750 | \nu = 30) = 0.005$

$P(-6.314 < T < 6.314 | \nu = 1) = 0.90 \leftarrow$ Symmetry saves work!

$P(-2.750 < T < 2.750 | \nu = 30) = 0.99 \leftarrow$ Symmetry saves work!

What happens as ν increases?

Summary: Student t Distribution

- Symmetric and "mound-shaped" distribution.
- Values near mean (which is 0) are more likely.
- One parameter, ν, or degrees of freedom.
- Unbounded support, $(-\infty, \infty)$
- To find probabilities, use Student t table.

Robustness of t Statistic

- Strictly speaking, the derivation of t distribution depended on the assumption that the sample was from a normal population. In practice, confidence intervals based on the t distribution work reasonably well, even when the population distribution is only approximately mound-shaped.

- This feature of t statistic is called "robustness".

Summary

Sample size	σ is known	σ is unknown
$n < 30$	Use t-statistic Use critical values of t distribution	Use t-statistic Use critical values of t distribution
$n > 30$	Use z-statistic Use critical values of standard normal distribution	Use t-statistic Use critical values of t distribution