# Introductory Review

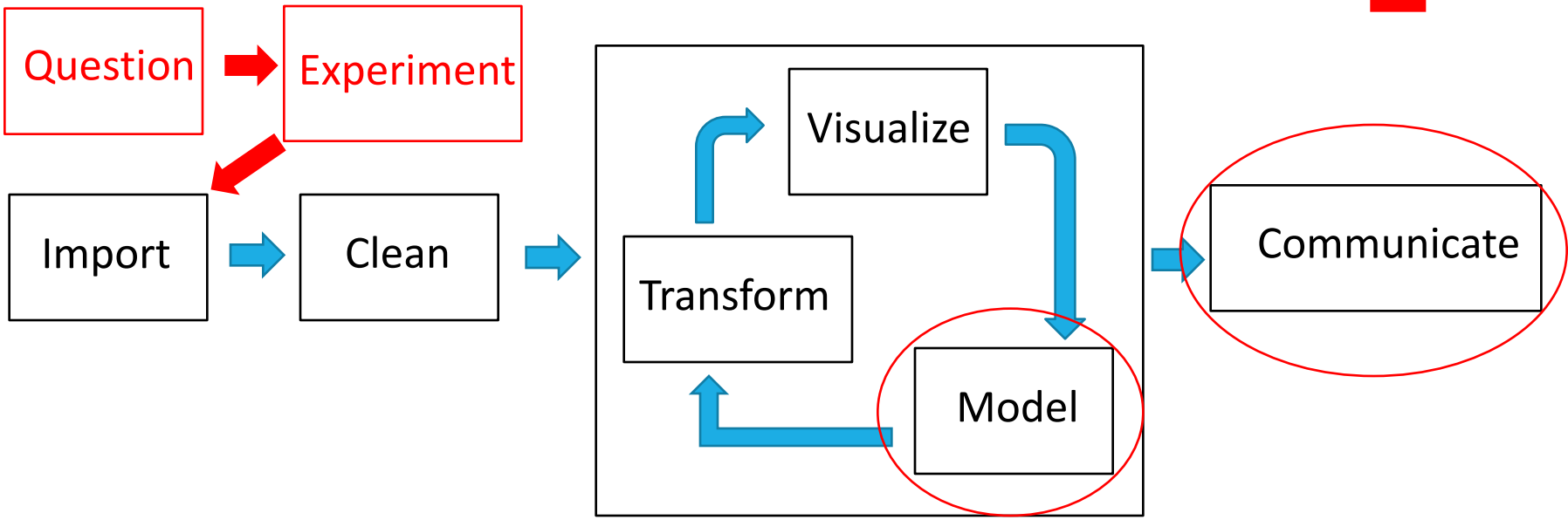
SAS

MEASURES OF CENTRALITY

CONFIDENCE INTERVALS

HYPOTHESIS TESTING

# Statistical flow chart




- Most statistics classes emphasize “Model” and “Communicate”
- It is important to keep in mind the entire process
- Actually, the flow chart is really a loop

# SAS

# Getting Started with SAS

SAS is a powerful tool for analyzing data

To get the most out of this class, you should become conversant in SAS

You can access SAS in at least three ways:

- A local copy installed on your computer
- A remote copy hosted by the “Citrix” SMU server
- A remote copy hosted by SAS (SAS onDemand)

# Remote Copy

https://odamid.oda.sas.com/SASODARegistration/

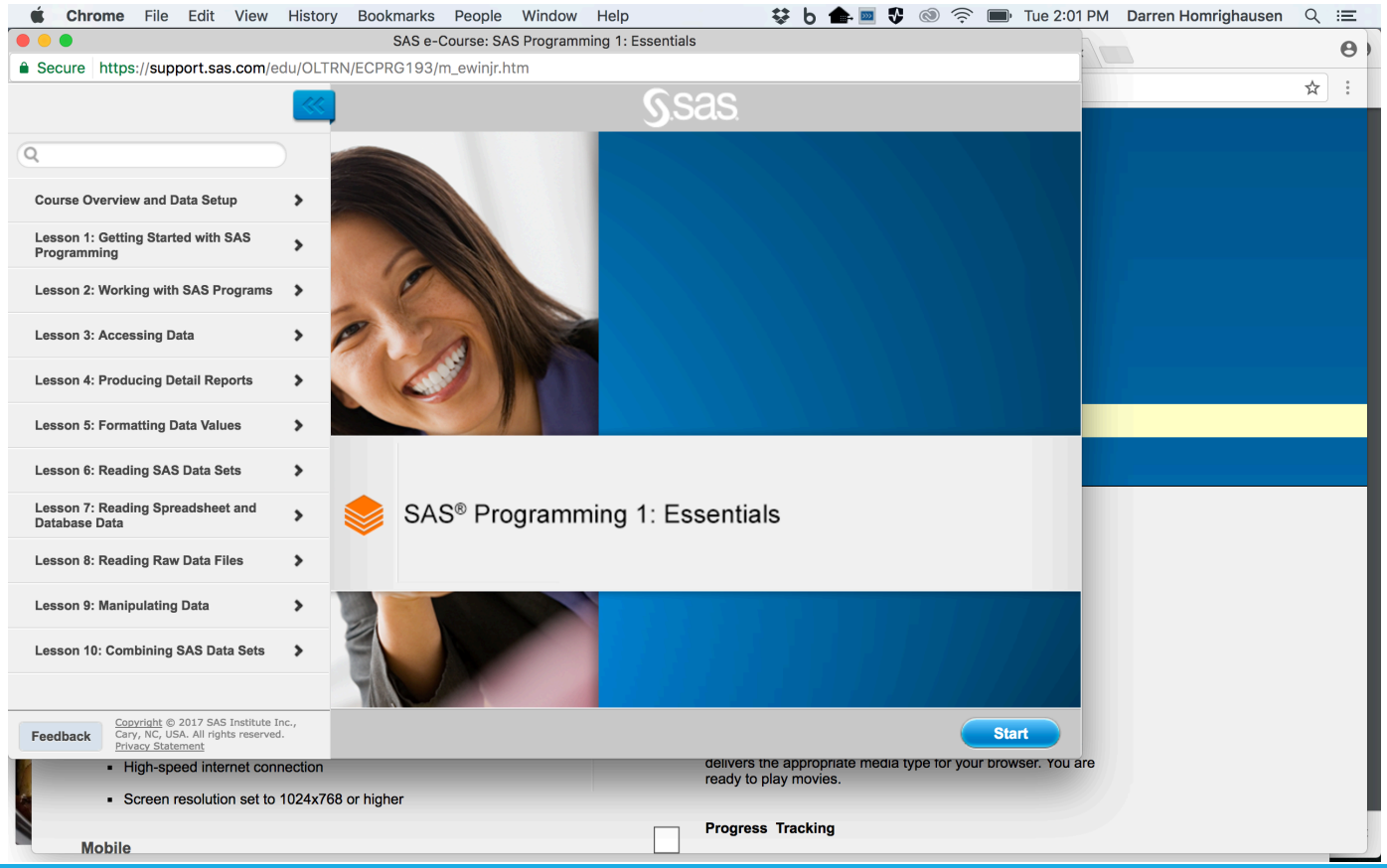
A confirmation email will be sent to the email address

After registering, you can join this class:

https://odamid.oda.sas.com/SASODAControlCenter/enroll.html?enroll=79763c48-3d66-4faa-b802-14db272bc718

# Learning SAS

After registering for SAS onDemand, you get access to some SAS tutorials:



# Learning SAS

Also, here are some links to brief introductory videos:

http://video.sas.com/detail/videos/programming/video/4573016765001/writing-a-basic-sas-program?autoStart=true

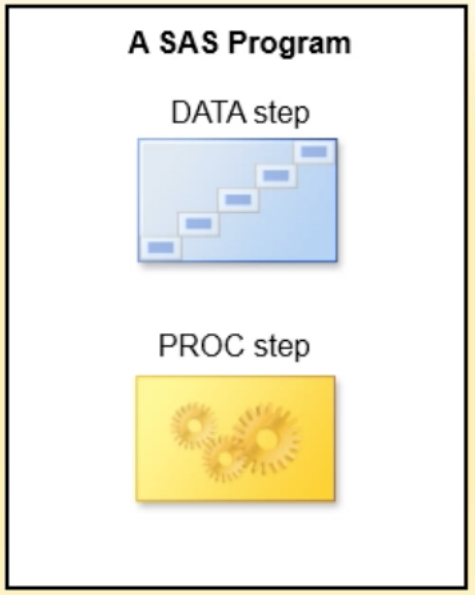
http://video.sas.com/detail/videos/managing-data/video/4573016761001/filtering-a-sas-table-in-a-data-step?autoStart=true

http://video.sas.com/detail/videos/managing-data/video/4573016759001/performing-conditional-logic-in-sas?autoStart=true

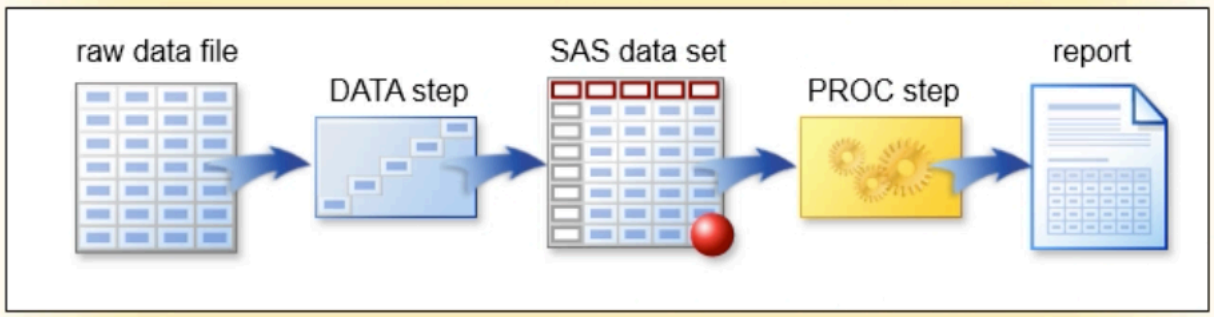
What does a typical SAS session look like?

# SAS overview

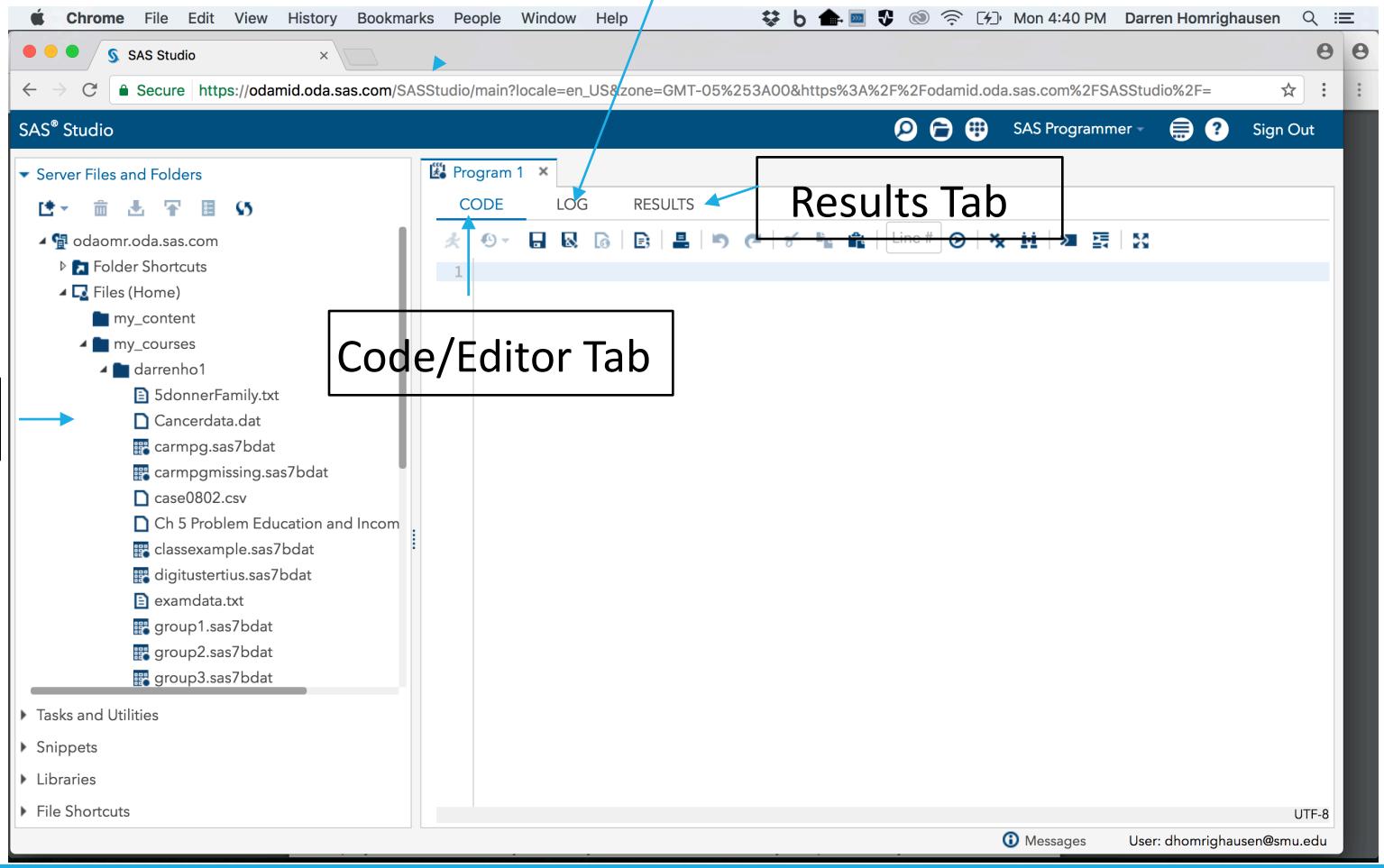
Every SAS session is some combination of DATA and PROC steps



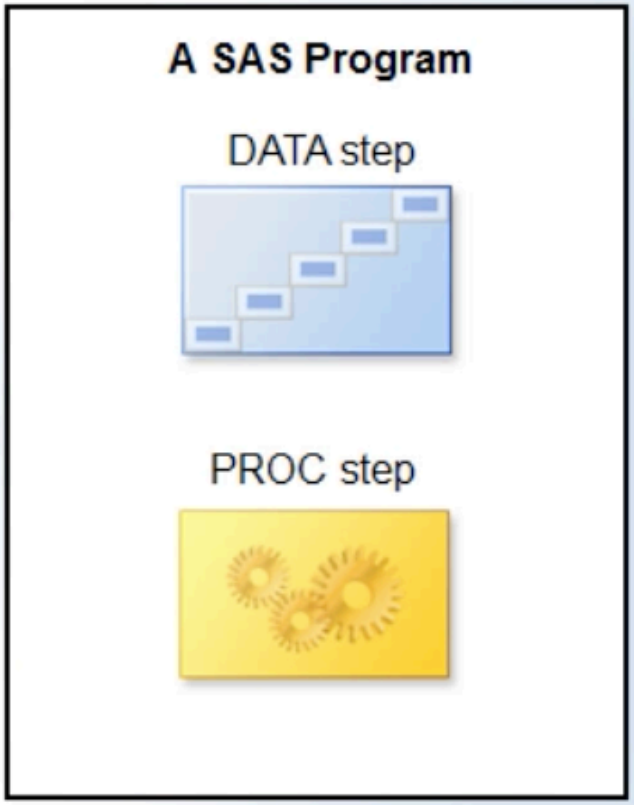

An example SAS session might look like:



# SAS session

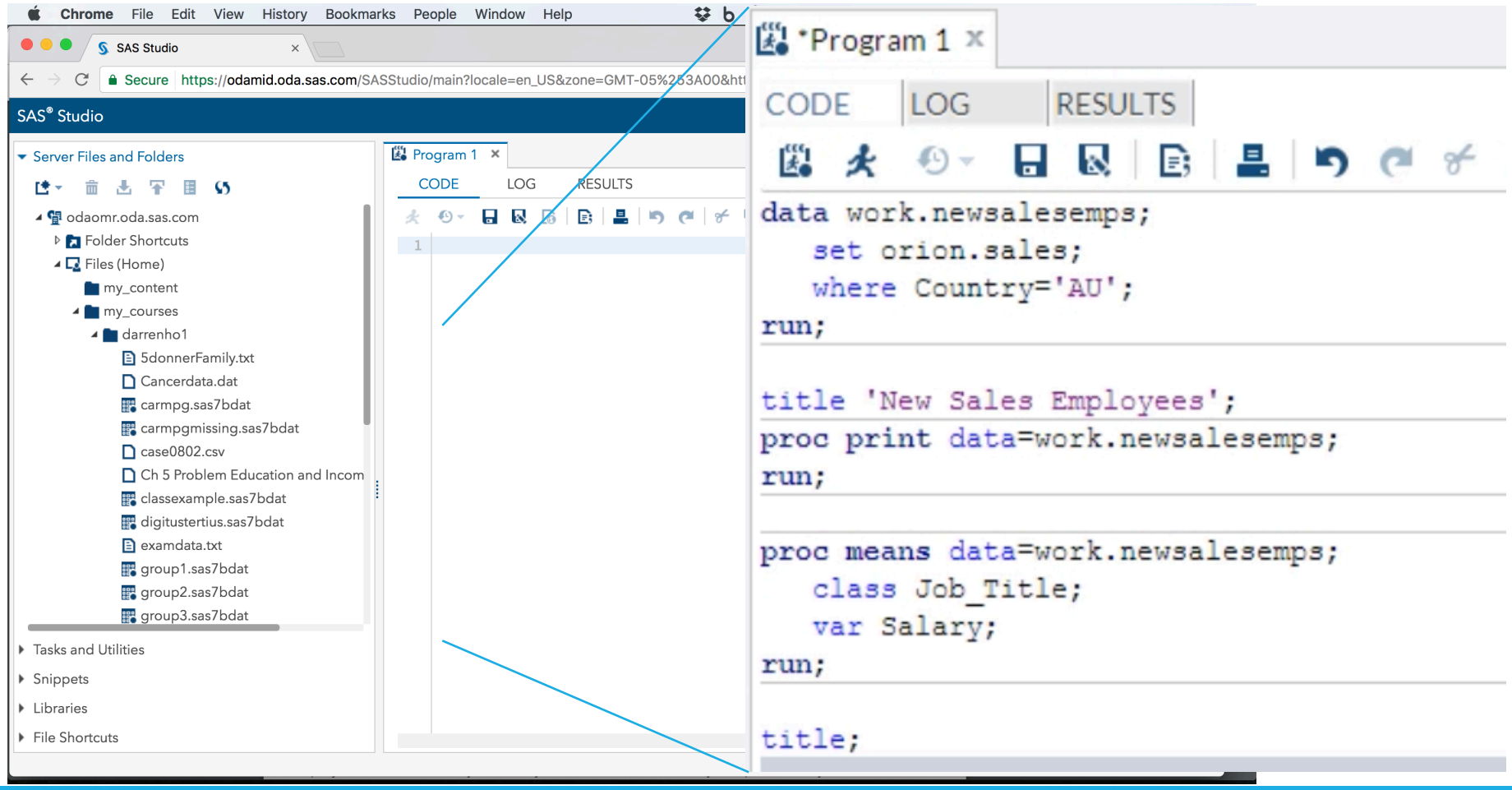

Log Tab
Results Tab
Code/Editor Tab
Files/Data
Chrome File Edit View History Bookmarks People Window Help
Mon 4:40 PM
Darren Homrighausen
SAS Studio
Secure https://odamid.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-05%253A00&https%3A%2F%2Fodamid.oda.sas.com%2FSASStudio%2F=
SAS® Studio
SAS Programmer
Sign Out
Server Files and Folders
odaomr.oda.sas.com
Folder Shortcuts
Files (Home)
my_content
my_courses
darrenho1
5donnerFamily.txt
Cancerdata.dat
carmpg.sas7bdat
carmpgmissing.sas7bdat
case0802.csv
Ch 5 Problem Education and Incom
classexample.sas7bdat
digitustertius.sas7bdat
examdata.txt
group1.sas7bdat
group2.sas7bdat
group3.sas7bdat
Tasks and Utilities
Snippets
Libraries
File Shortcuts
Program 1
CODE
LOG
RESULTS
UTF-8
Messages
User: dhomrighausen@smu.edu

# SAS session: Editor

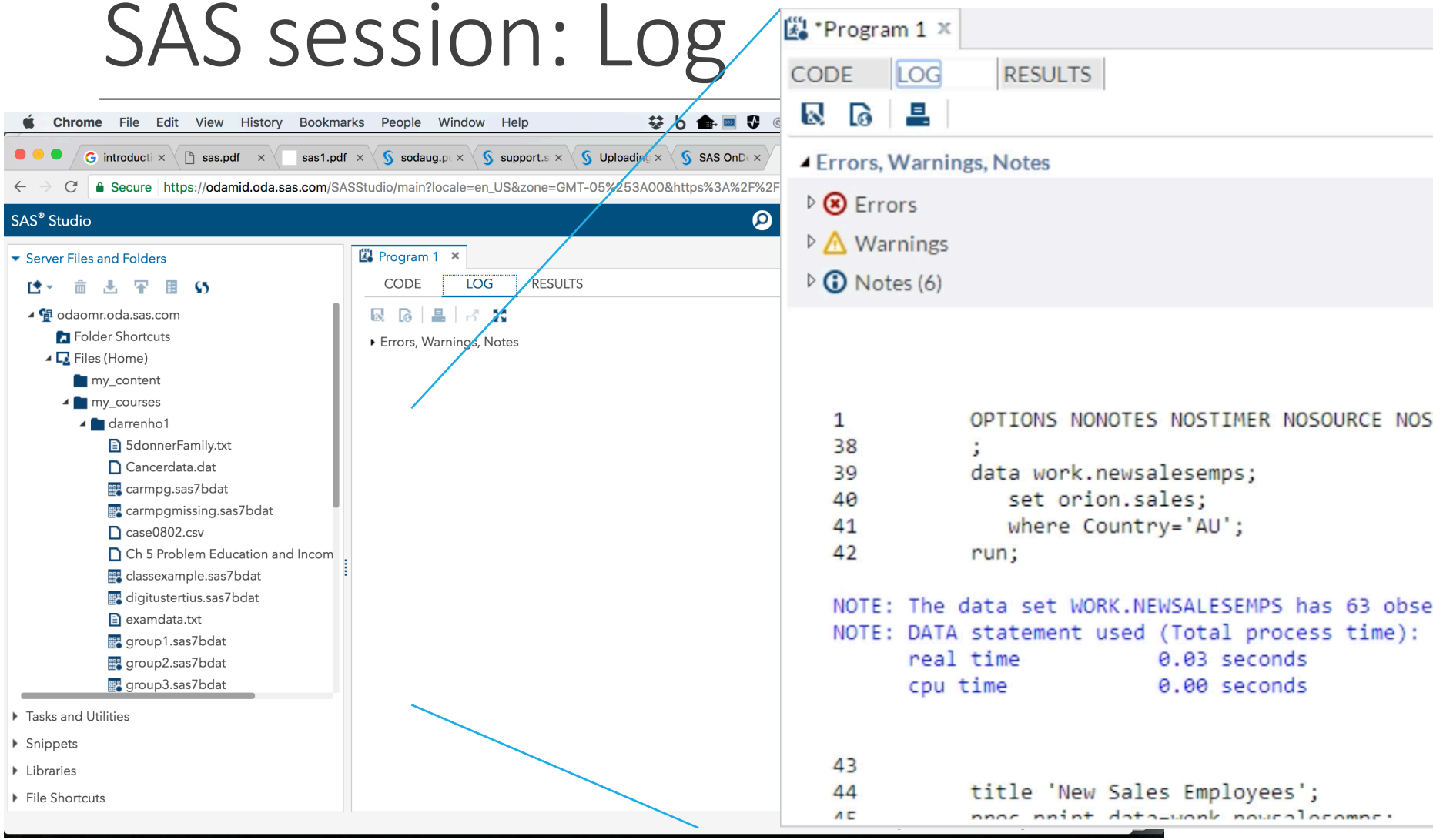



```
data work.newsalesemps;
   set orion.sales;
   where Country='AU';
run;

proc print data=work.newsalesemps;
run;

proc means data=work.newsalesemps;
   class Job_Title;
   var Salary;
run;
```

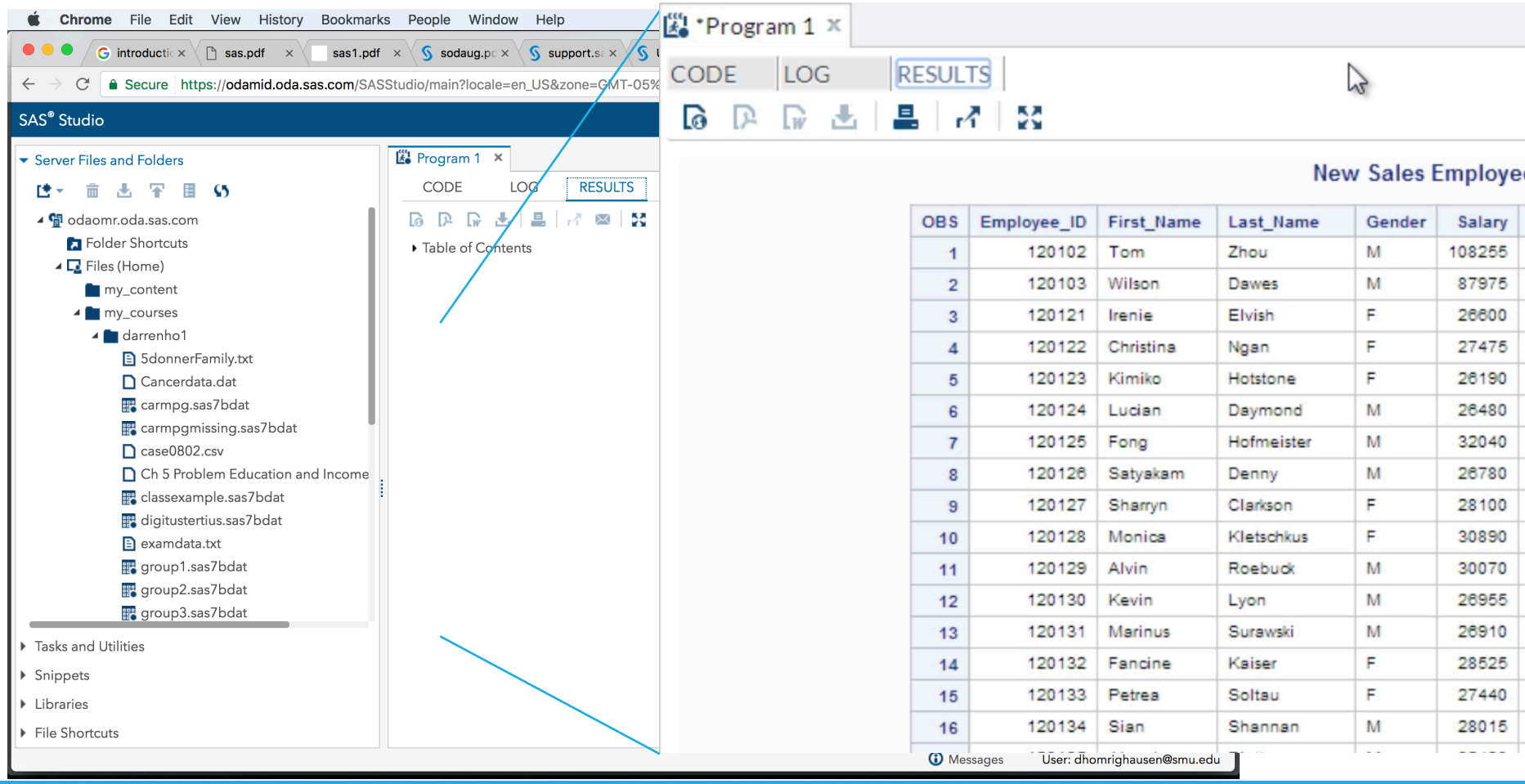
# SAS session: Editor/Code



# SAS session: Log

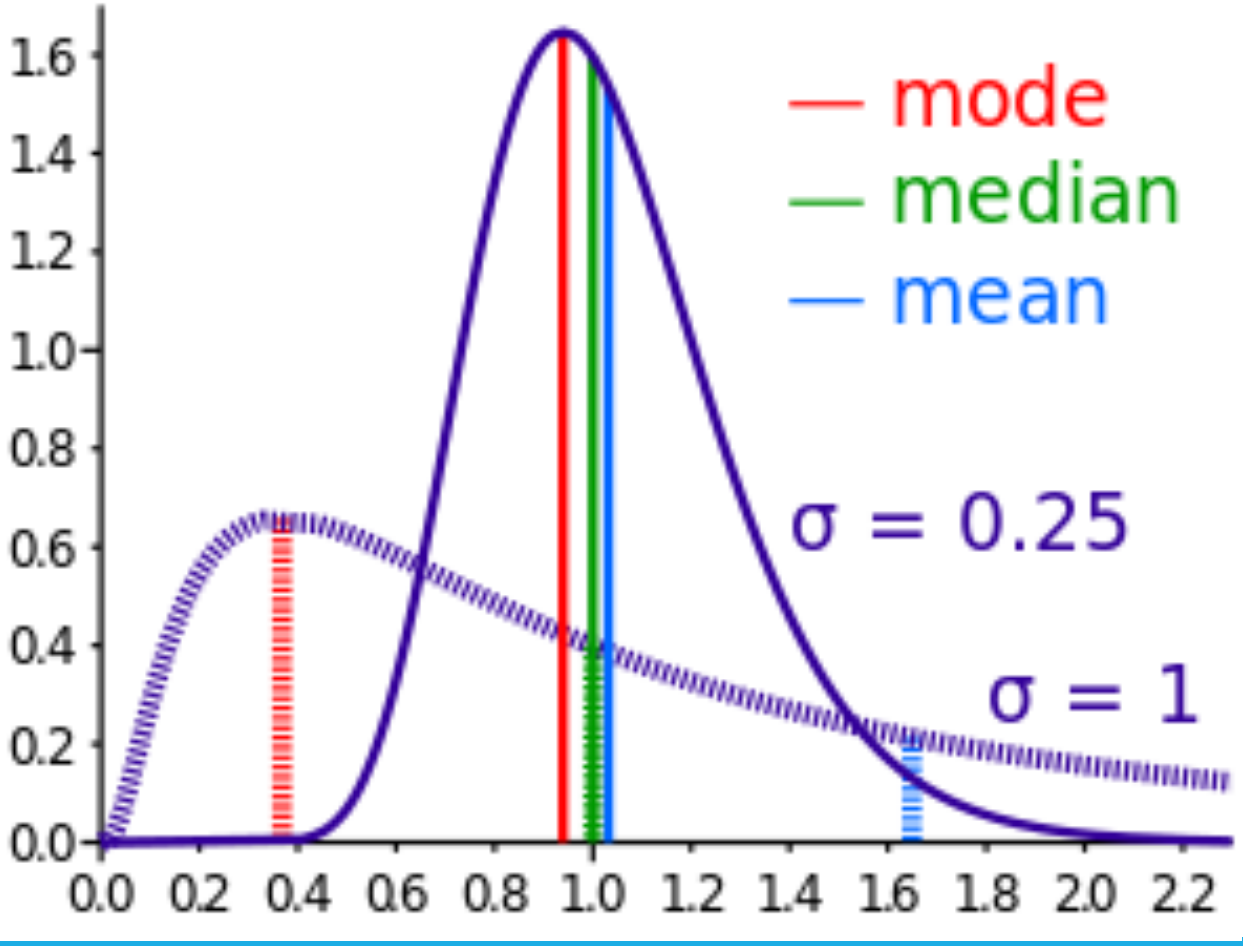


# SAS session: Results



# Measures of centrality

# Mean, Median, and Mode



# Estimating the population mean

# Assumptions

We will return to these later…

# The sample mean

If $Y_1, Y_2, \ldots, Y_n$ are the possible values and $y_1, y_2, \ldots, y_n$ are observed, then

$$\bar{Y} = \frac{(Y_1 + Y_2 + \ldots + Y_n)}{n} \text{ and } \bar{y} = \frac{(y_1 + y_2 + \ldots + y_n)}{n}$$

The idea: $\bar{Y}$ "approx. equals" $\mu$

What is the definition of "approx. equals"?

- The mean of $\bar{Y}$ equals $\mu$
- $\text{Variance}(\bar{Y}) = \frac{\sigma^2}{n} \rightarrow \text{SD}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$ (*note: this goes to 0 as n increases and is the law of large numbers*)

So, $\bar{Y} = \mu + \text{remainder}$,

where $|\text{remainder}|$ is "unlikely" to exceed $(\text{multiple}) \cdot \frac{\sigma}{\sqrt{n}}$

How do we quantify "unlikely"? Via the central limit theorem (CLT)!

# The central limit theorem CLT)

The CLT is a refinement of the law of large numbers

It states that not only does $\bar{Y}$ eventually converge to $\mu$, but the distribution of $\bar{Y}$ "eventually" looks like normal/Gaussian distribution

("*eventually": commonly taken to be n>30 and obs. are independent*)

In notation, we write this as $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$

So, $\bar{Y} = \mu + \text{remainder} \leftrightarrow \bar{Y} - \mu = \text{remainder}$

Hence, $\text{remainder} \sim N(0, \frac{\sigma^2}{n})$

# Confidence interval for the population mean

Suppose we want to estimate $\mu$

A reasonable estimator is $\hat{\mu} = \bar{Y}$

What if we additionally want to find an interval which we believe contains $\mu$?

If that interval is symmetric around $\bar{Y}$, it would look like $[\bar{Y} - E, \bar{Y} + E]$

Our task: Quantify how "likely" it is that: $\bar{Y} - E \leq \mu \leq \bar{Y} + E$

For that, we use the result: $\text{remainder} \sim N(0, \frac{\sigma^2}{n})$

Hence, choose: $E$ = $(\text{multiple}) \cdot \frac{\sigma}{\sqrt{n}}$

where $(\text{multiple})$ is chosen to quantify "likely"…

# Confidence interval for the population mean

Reminder:

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$ , $\bar{Y} = \mu + \text{remainder}$, $\text{remainder} \sim N\left(0, \frac{\sigma^2}{n}\right)$, $E$ = (multiple) $\cdot \frac{\sigma}{\sqrt{n}}$

→ set (multiple) to be a quantile of the standard normal/Gaussian distribution

(*we notate the normal distribution as* $Z \sim N(0,1)$)

$z_\alpha$ is the $\alpha^{th}$ quantile of a normal provided: $\text{Probability}(Z < z_\alpha) = 1 - \alpha$

Claim: If we set $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ then $[\bar{y} - E, \bar{y} + E]$ is a $(1-\alpha)100\%$ confidence interval (CI)

(Proof sketch)

$$\text{Probability}(\bar{Y} - E \leq \mu \leq \bar{Y} + E) = \text{Probability}(-E \leq \text{remainder} \leq E)$$

$$= \text{Probability}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

# Example:

Claim: If we set $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ then $[\bar{y} - E, \bar{y} + E]$ is a $(1-\alpha)100\%$ confidence interval (CI) for $\mu$

E is often referred to as the margin of error

Find a 99% confidence interval for $\mu$ if $\sigma^2 = 1.44$ and a sample size of 49

$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 2.575 \cdot \frac{1.2}{\sqrt{49}} = 0.4414$ (note: this isn't a CI)

# Hypothesis testing

- THE POLYCEPHALIC TWIN OF CONFIDENCE INTERVALS

# Let's look at the mean

- The mean $\mu$ is unknown
- Suppose we presume $\mu = \mu_0$ (say, $\mu_0$ = 10)
- We calculate $\bar{y}$ (say, $\bar{y}$ = 13)
- Is this sufficient evidence that $\mu \neq \mu_0$?

How likely is this observed value? Can we say "beyond a reasonable doubt" that this observed value is too extreme?

# How much variation should we expect?

The mean of $\bar{Y}$ is $\mu$

The Variance$(\bar{Y}) = \frac{\sigma^2}{n} \rightarrow$ SD$(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$

Therefore, if we center and rescale:

$\rightarrow Z = \frac{\bar{Y}-\mu}{(\sigma/\sqrt{n})}$ has mean 0 and variance 1, by CLT: $Z \sim N(0,1)$
(*assuming $n$ is large enough and obs. are independent*)

Large values of $|Z|$ are "unlikely": Probability$(|Z| > z_{\alpha/2}) = \alpha$

Of course, computing this $Z$ requires knowledge of $\mu$ which is unknown

What we do know is the hypothesized value $\mu_0$

# How much variation should we expect?

Let's look at $Z_0 = \frac{\bar{Y} - \mu_0}{(\sigma/\sqrt{n})}$ and compute $z_0 = \frac{\bar{y} - \mu_0}{(\sigma/\sqrt{n})}$ instead

As before, computing $\text{Probability}(|Z_0| > z_0)$ requires knowledge of $\mu$
(*after all, we showed that the distribution of $\bar{Y}$ depends on $\mu$*)

Instead, we compute a related quantity: $\text{Probability}(|Z_0| > z_0 \,|\, \mu = \mu_0)$
(*this is a conditional probability under the hypothesis that $\mu = \mu_0$*)

<u>Important:</u> If $\mu = \mu_0$, then we of course know $\mu$ as we know $\mu_0$

Hence, we can compute $\text{Probability}(|Z_0| > z_0 \,|\, \mu = \mu_0)$ using the CLT:

$\rightarrow Z_0 \,|\, \mu = \mu_0$ is approximately $N(0,1)$

# Hypothesis Testing: Definitions

By convention, we define the following key ideas

- $H_0$: $\mu = \mu_0$ a "null hypothesis"
- $H_a$: $\mu \neq \mu_0$ an "alternative hypothesis"
- $\text{Probability}(|Z_0| > z_0 | \mu = \mu_0)$ as the "p-value"

(*sometimes alternatively written:* $Probability(|Z_0| > z_0 | H_0)$)

**Massively important:** the p-value is not $\text{Probability}(\mu = \mu_0 \mid |Z_0| > z_0)$

Implication: the p-value…

…is not the conditional probability of the null hypothesis given $|Z_0| > z_0$

…is the conditional probability of getting a larger observation given $H_0$

If this probability is "small enough", we have evidence to reject $H_0$

# Hypothesis testing: An analogy

Suppose we are on a jury at a trial

Our verdict will either be guilty or innocent after seeing evidence

The doctrine of "innocent until proven guilty" demands evidence in excess of "reasonable doubt"

Statistics allows us to quantify both "innocent until proven guilty" and "beyond reasonable doubt" for a wide array of scientific problems

- "innocent until proven guilty" ↔ hypothesis testing
- "innocence" ↔ null hypothesis
- "guilty" ↔ alternative hypothesis
- "beyond reasonable doubt" ↔ small p-value

# Confidence intervals and hypothesis tests

Suppose for a given problem we compute $z_0 = 1.96$

How likely is it that we would draw a $|Z_0| > z_0$, given $\mu = \mu_0$?

→ Probability$((|Z_0| > z_0 | \mu = \mu_0) = 0.05$

Let's compare this to a (1- $\alpha$)$100\% = 95\%$ confidence interval

Reminder:

$z_\alpha$ is the $\alpha^{th}$ quantile of a normal provided: Probability$(Z < z_\alpha) = 1 - \alpha$

→ $0.025^{th}$ quantile is the value $z_{0.025}$ such that
Probability$(|Z| > z_{0.025}) = 0.05$

When writing $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ the multiplier $z_{0.025} = 1.96$, a main ingredient for a 95% confidence interval

# The connection

$100(1-\alpha)\%$ confidence intervals will include $\mu_0$
if and only if
Probability$((|Z_0| > z_0 | \mu = \mu_0) \geq \alpha$

<u>Stated another way:</u>

$\mu_0$ not being in a $100(1-\alpha)\%$ confidence interval is equivalent to rejecting $H_0$: $\mu = \mu_0$ with a p-value less than $\alpha$

(*There are some caveats here. If Z takes on too few of values or if we are doing "one sided hypothesis tests", some adjustments are necessary*)