



# Making Data Useful

IT Professionals Conference, June 22 2017

Russell Dimond

Associate Director/Statistical Consultant

Social Science Computing Cooperative



# How will these be used?

1. On a scale of 1-5 how would you rate the IT Professionals Conference?
2. Will you attend this conference next year?
3. Would you recommend this conference to your colleagues?
4. What went well with the conference?
5. What could we improve next year?



# How will these be used?

1. How long was it before you got an initial response from a staff person regarding your request?
2. Was this an acceptable response time?
3. How satisfied were you with the service you received?
4. Did the staff person conduct him/herself in a completely professional and courteous manner?

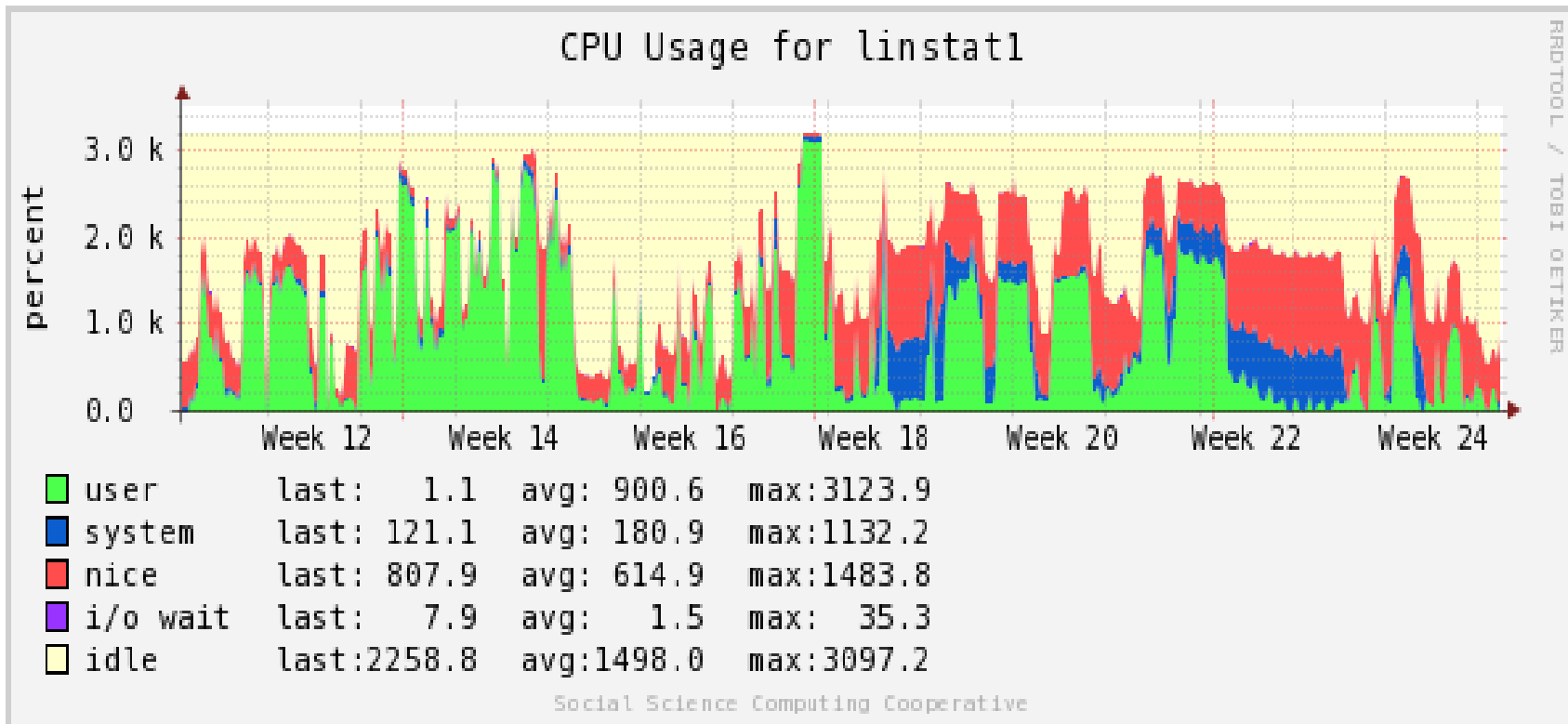


# Does SSCC need more servers?

What are we trying to avoid?

What data do we need?

How should we analyze it?

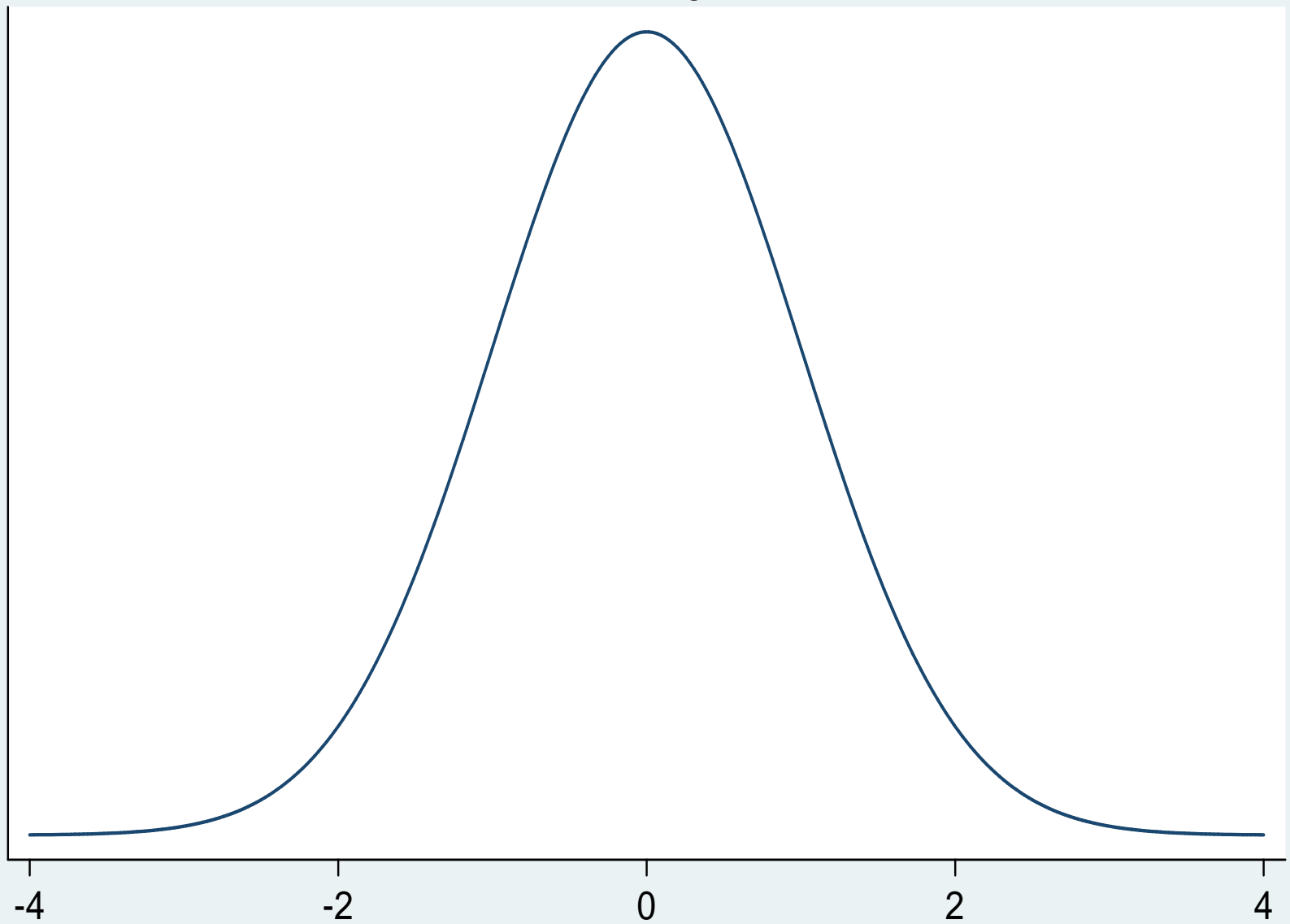


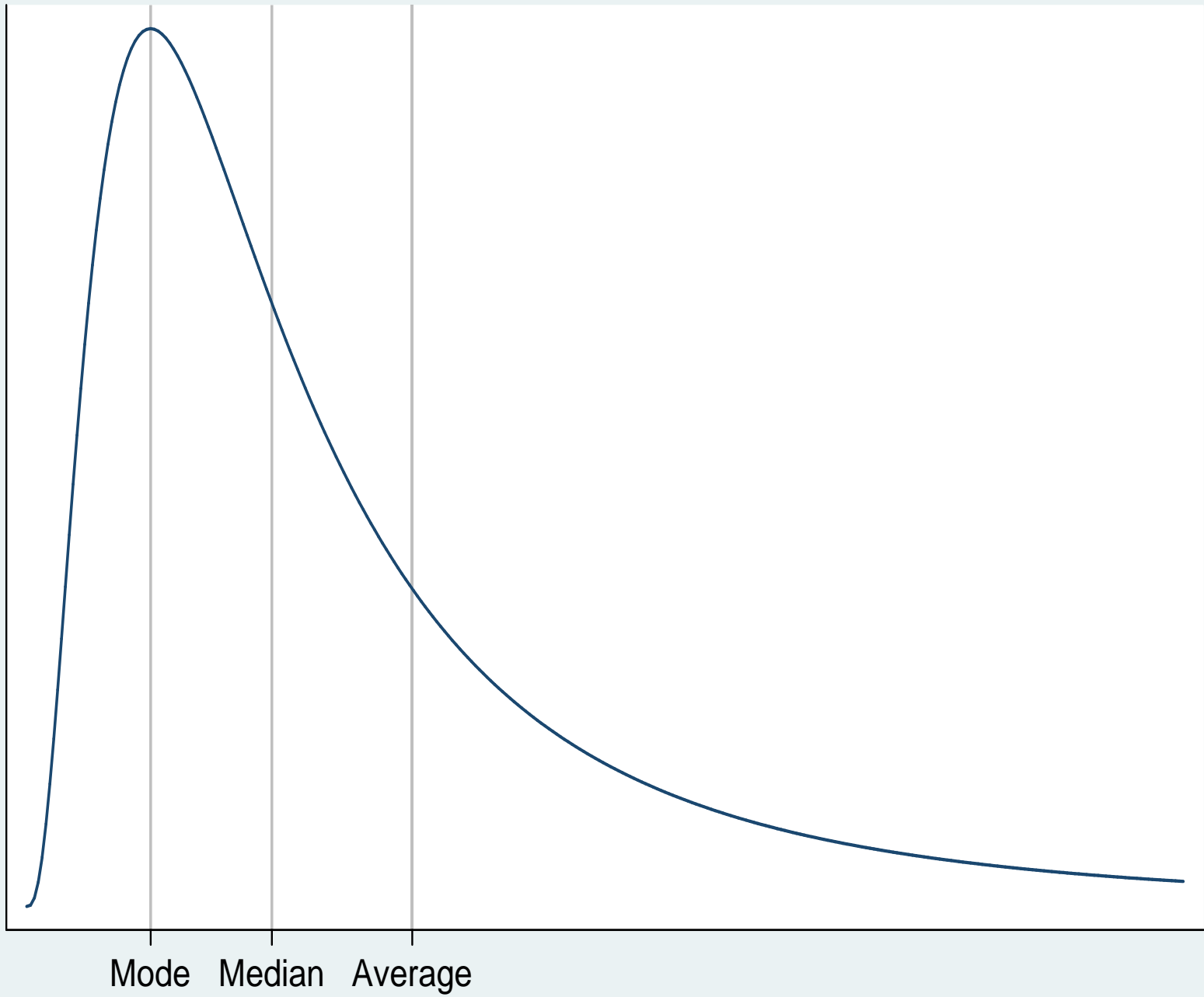


# Tools for Understanding Data

# "Normal" Distribution

Mean=0, Sigma=1







# Tools for Understanding Data

- What's the variable? Winstat CPU time used (in hours)
- What's an observation? An SSCC member
- How many observations? 1,420
- Mean: 35.3
- Std. Dev.: 195.5





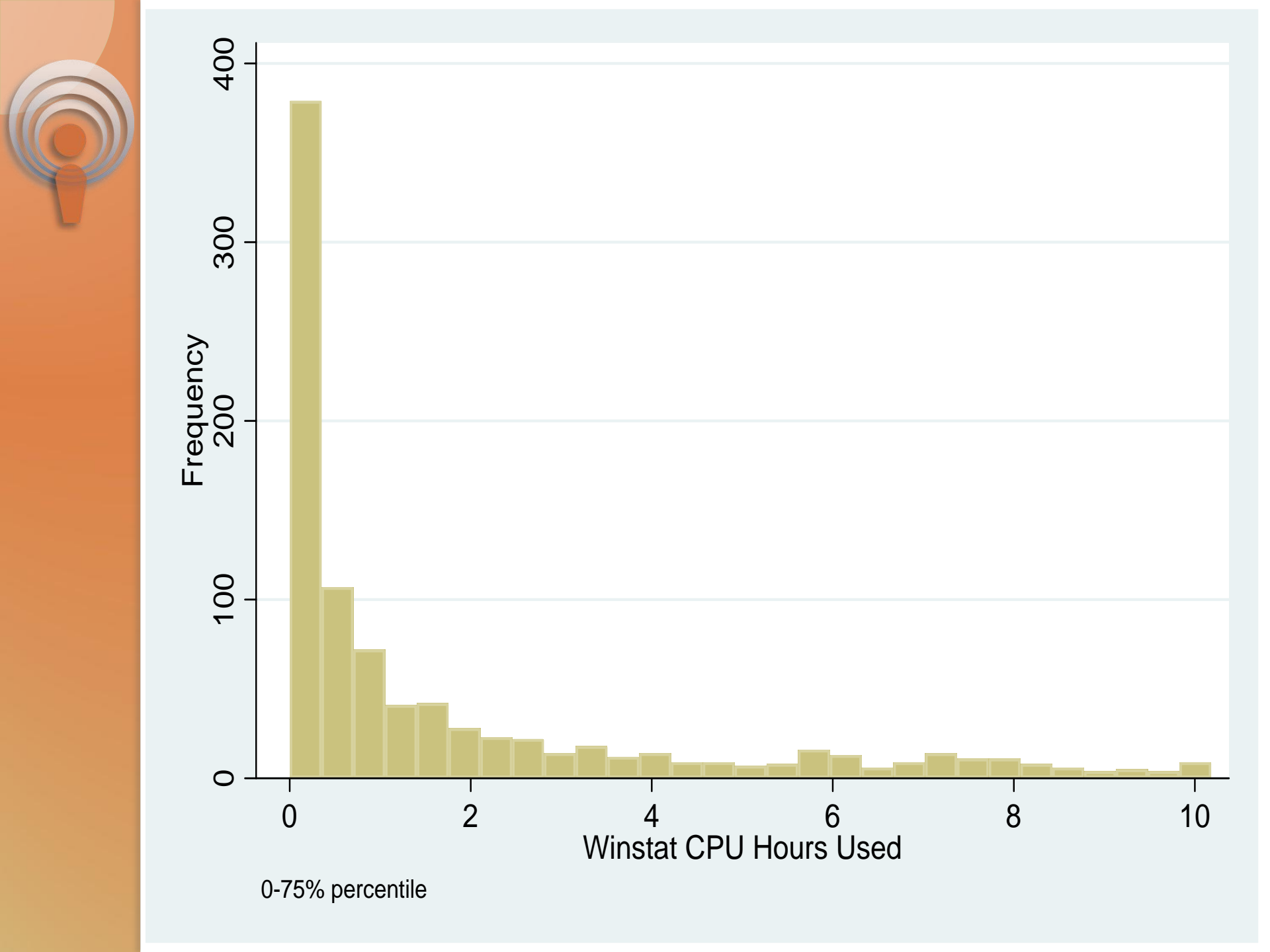
# Percentiles (Five Number Summary)

- Min: 0
- 25%: 0.04
- 50% (Median): 0.9
- 75%: 7.7
- Max: 3,845



# Exclude Non-Users

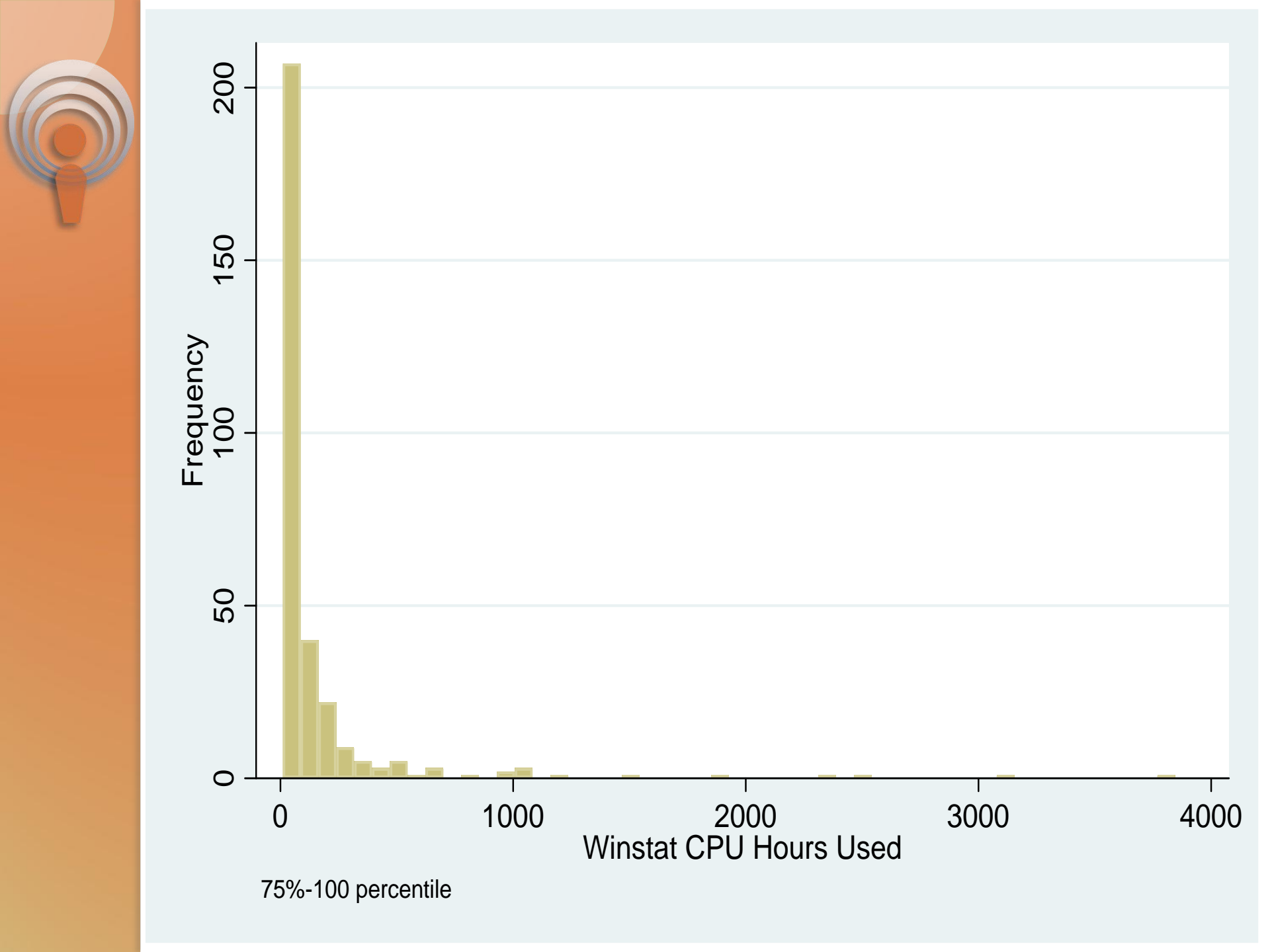
- 86.5% of members used Winstat
- Mean: 40.7
- Std. Dev.: 209.7
- Min: 0.0005
- 25%: 0.2
- 50% (Median): 1.5
- 75%: 10.3
- Max: 3,845





# Focusing on the High End

- 75%: 10.3
- 90%: 65.9
- 95%: 162.7
- 99%: 814.6
- 1,911
- 2,375
- 2,475
- 3,145
- 3,845



Frequency

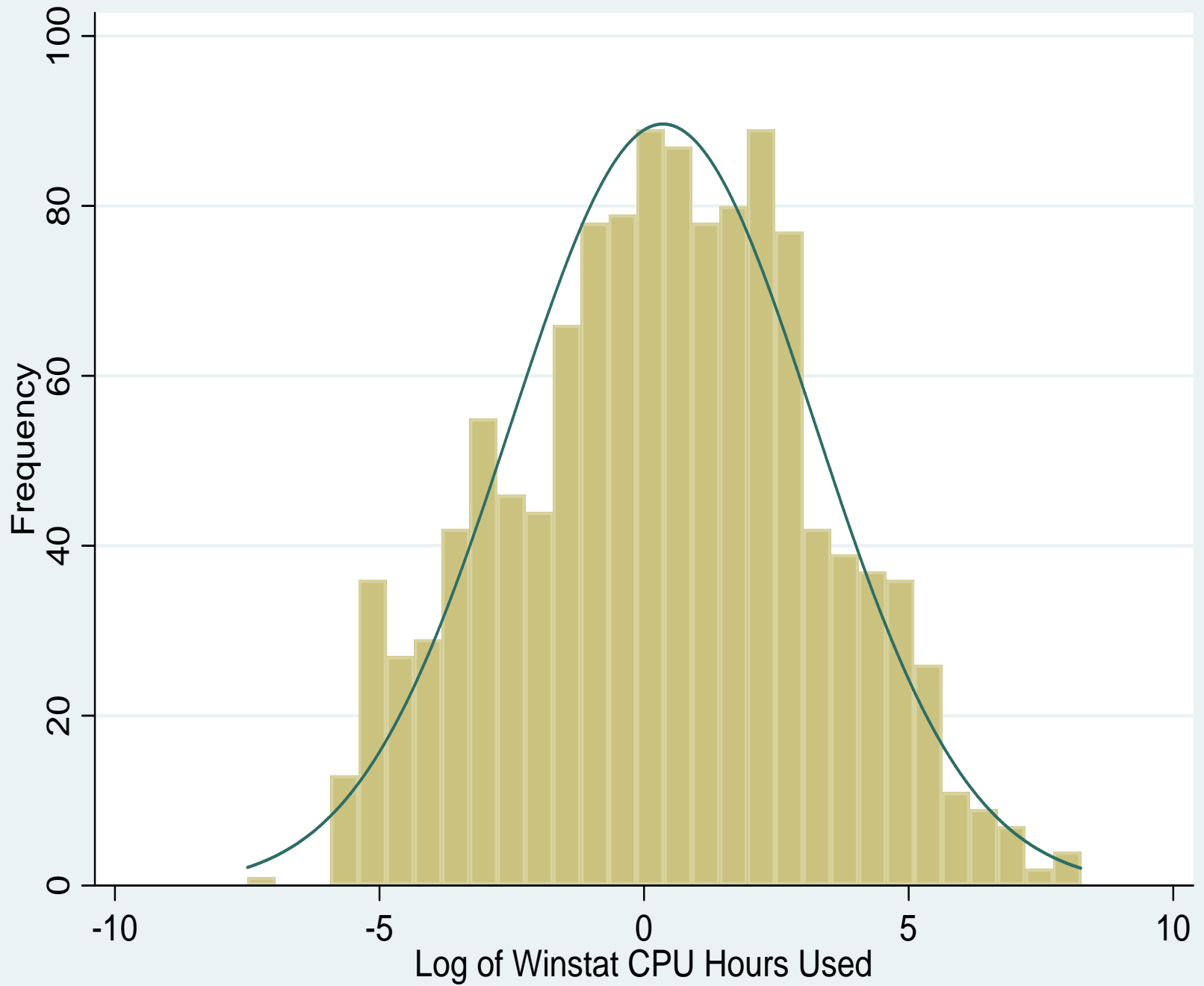
Winstat CPU Hours Used

75%-100 percentile



# What did we learn?

- Some members (not many) don't use Winstat at all
- Most members use little CPU time
- A few use a lot
- Extremely high-end usage is not an anomaly: it follows a consistent pattern we can expect to persist





# Types of Data

- Bounded (e.g. minimum of zero)
- Interval (min and max)
- Count
- Ordered Categorical
- Unordered Categorical
- Binary





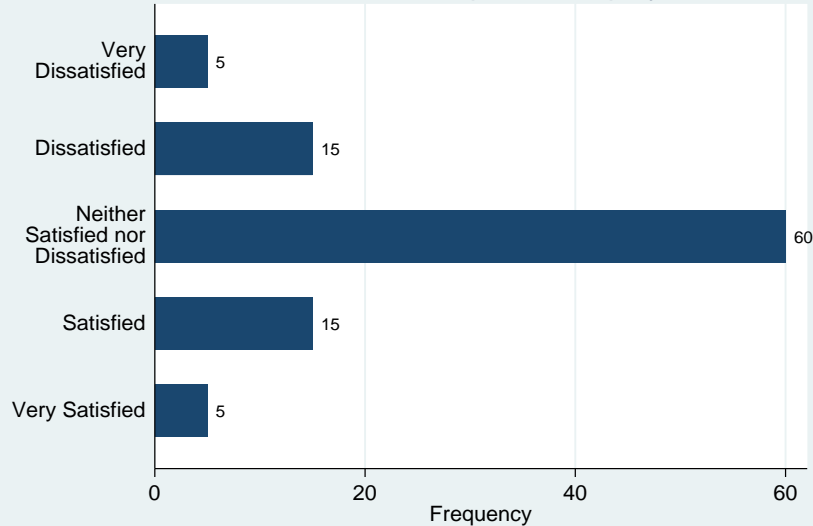
# Tool for Understanding Categorical Data: Frequencies

Satisfaction with Help From Employee One	Frequency
Very Dissatisfied	5
Dissatisfied	15
Neither Satisfied nor Dissatisfied	60
Satisfied	15
Very Satisfied	5

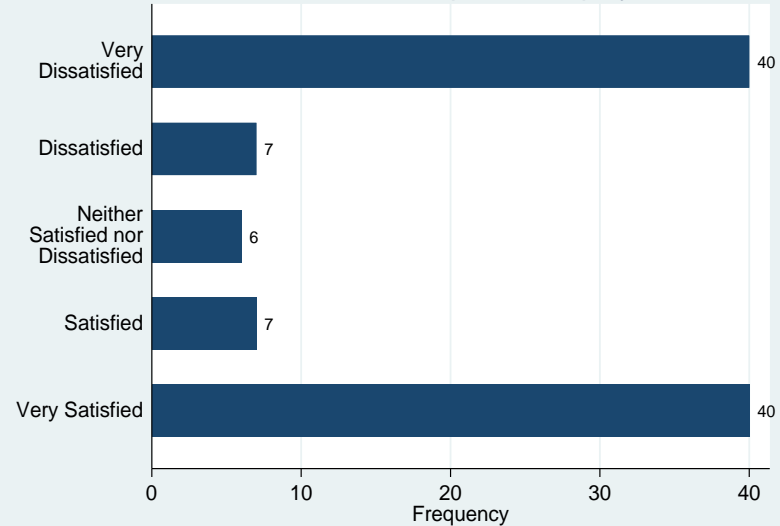
Satisfaction with Help From Employee Two	Frequency
Very Dissatisfied	40
Dissatisfied	7
Neither Satisfied nor Dissatisfied	6
Satisfied	7
Very Satisfied	40



Satisfaction with Help from Employee One



Satisfaction with Help from Employee Two



**How are they different?**  
**How would you help them?**



# Tool for *Misunderstanding* Categorical Data: Averages

- Average Satisfaction for Employee One: 3
- Average Satisfaction for Employee Two: 3

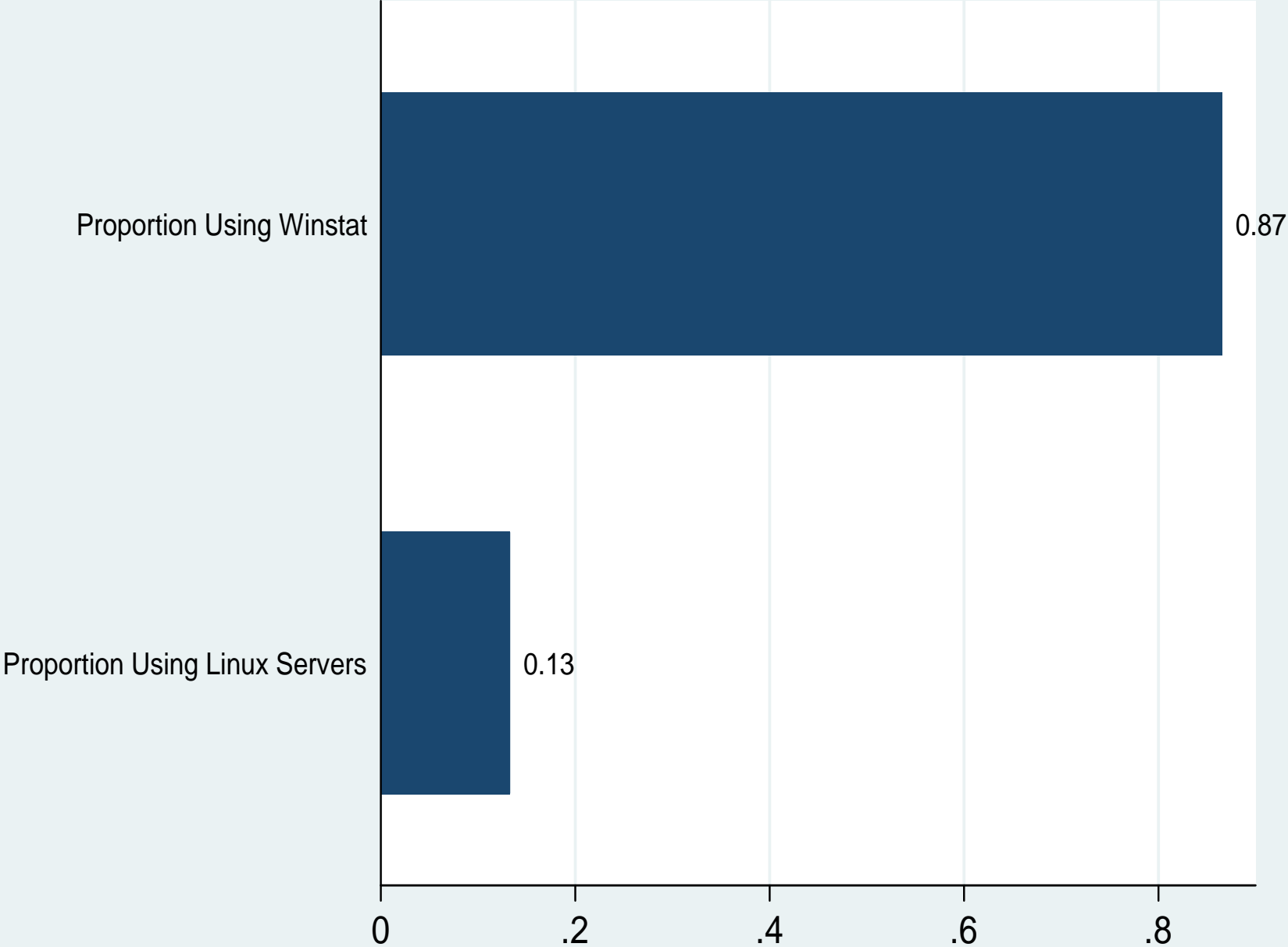


# Tools for Understanding Binary Data

- Frequencies & Bar Graphs
- Averages(!)



# Proportion of SSCC Members using Winstat and Linux Servers





# What to Remember...

- Before you start collecting any data, think carefully about how you'll use it
- Averages are useful for understanding normally distributed data
- Use percentiles and histograms to understand non-normal data
- Use frequencies and percentages to understand categorical data