WHY GOOD MODELS FAIL

HOW AMBIGUITY, UNCERTAINTY, AND BAD SCIENCE
CAUSE MOST DATA SCIENCE PROJECTS TO FAIL

87%

OF ALL DATA SCIENCE PROJECTS FAIL
-VENTUREBEAT

REASON ONE: **AMBIGUITY**

WHEN REQUIREMENTS ARE HARD TO DEFINE

AMBIGUITY OCCURS WHEN WE DON'T KNOW WHAT QUESTIONS TO ASK

HOW TO NAVIGATE **AMBIGUITY** IN DATA SCIENCE?

1. FILL KNOWLEDGE GAPS

BECOME A SUBJECT MATTER EXPERT

2. ITERATE QUICKLY

RESEARCH, PRESENT, GET FEEDBACK

3. DEFER TO THE EXPERTS

LET THEM DRIVE THE RESEARCH

4. EXPLORE WITHOUT BUILDING

UNDERSTAND THE PROBLEM FULLY, THEN START DESIGNING

REASON TWO: **UNCERTAINTY**

WHEN SUCCESS IS HARD TO DEFINE

HOW TO NAVIGATE **UNCERTAINTY** IN DATA SCIENCE?

1. UNDERSTAND THE PROCESS

THE FULL LIFECYCLE OF A RECORD

2. UNDERSTAND THE KPIS

THE SOURCE, TARGET, CALCULATION, STAKEHOLDERS

3. UNDERSTAND THE DOLLARS

SAVINGS/REVENUE CALCULATIONS MUST BE DONE METHODICALLY

THE **PROCESS** IMPACTS THE **KPIS** WHICH IMPACT THE **DOLLARS**

REASON THREE: **BAD SCIENCE**

WHEN VALUE IS MIS-MEASURED

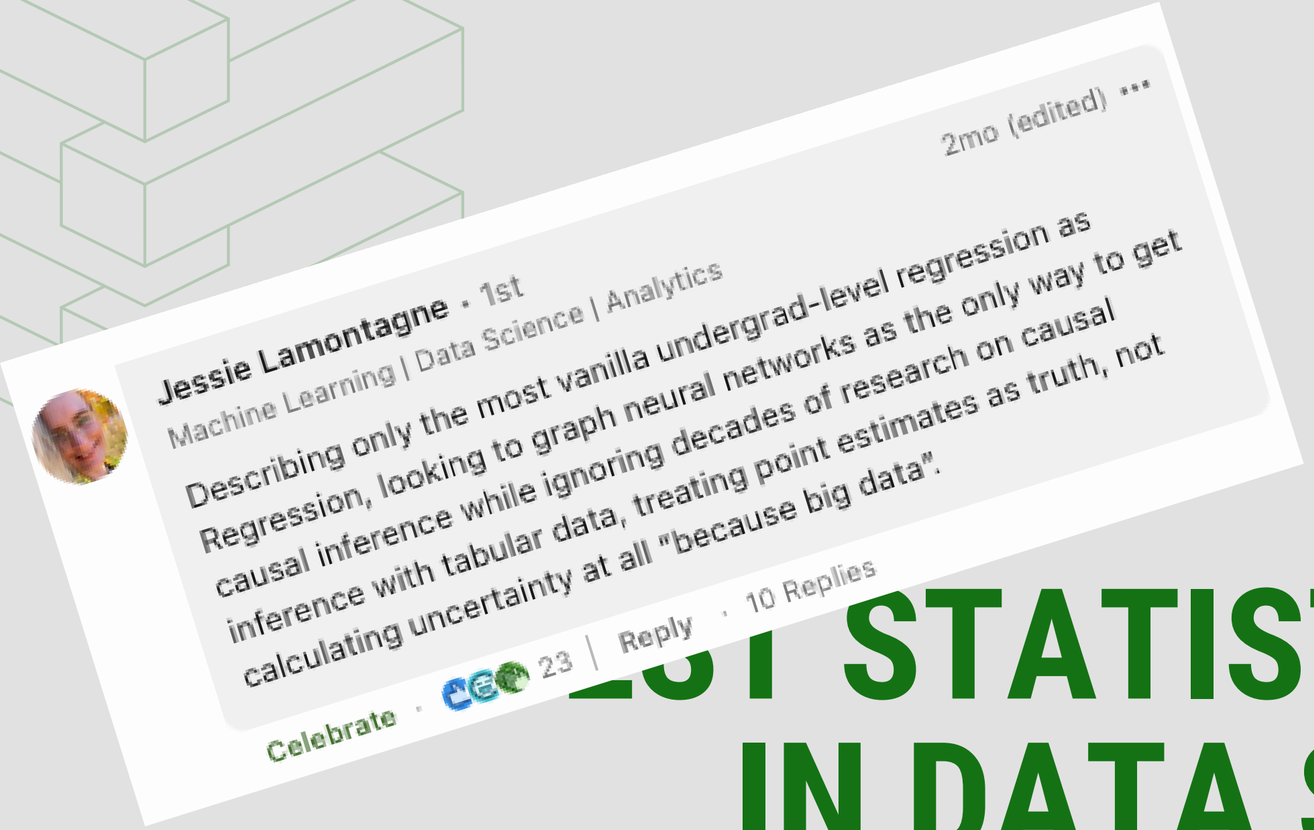
1. INCORRECT METRICS

INAPPROPRIATE METRICS ARE THE HALLMARK OF BAD SCIENCE

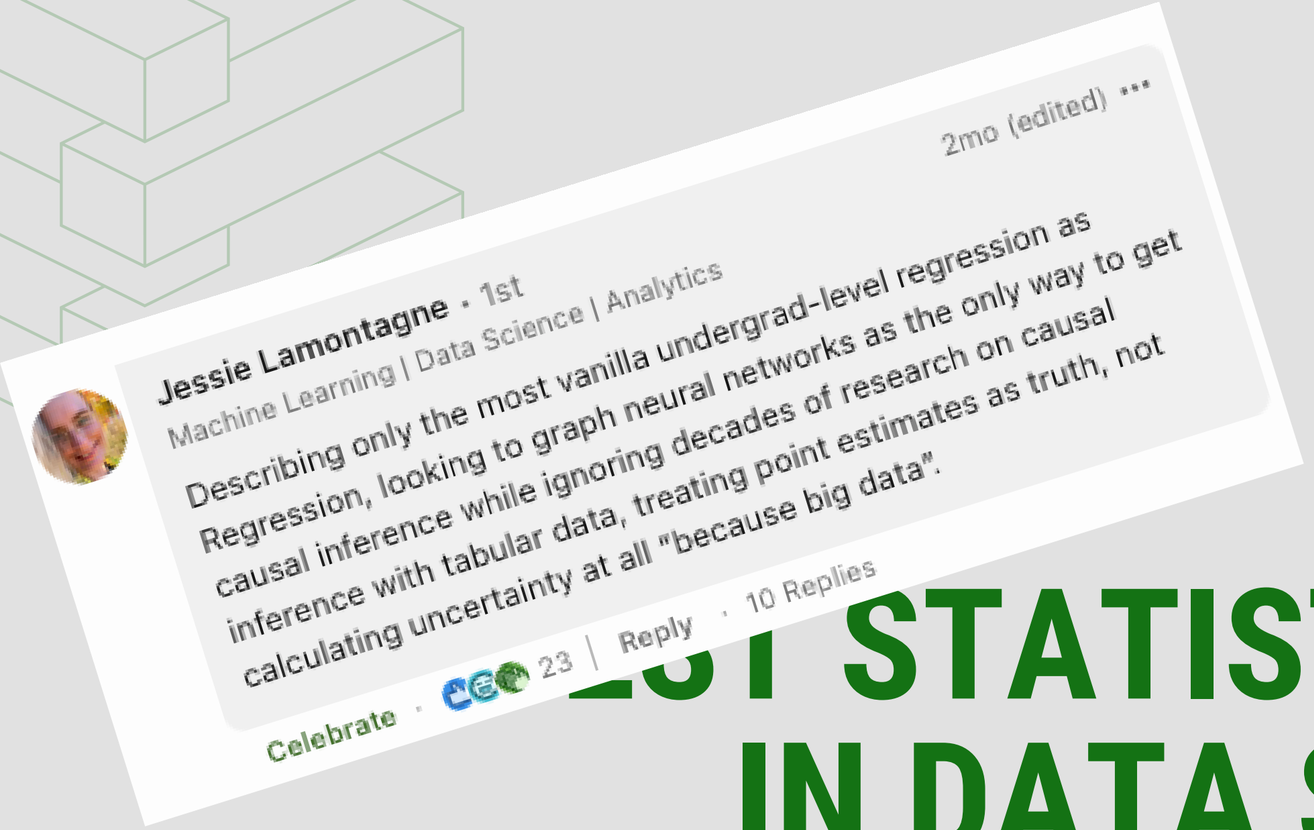
2. BAD STATISTICS

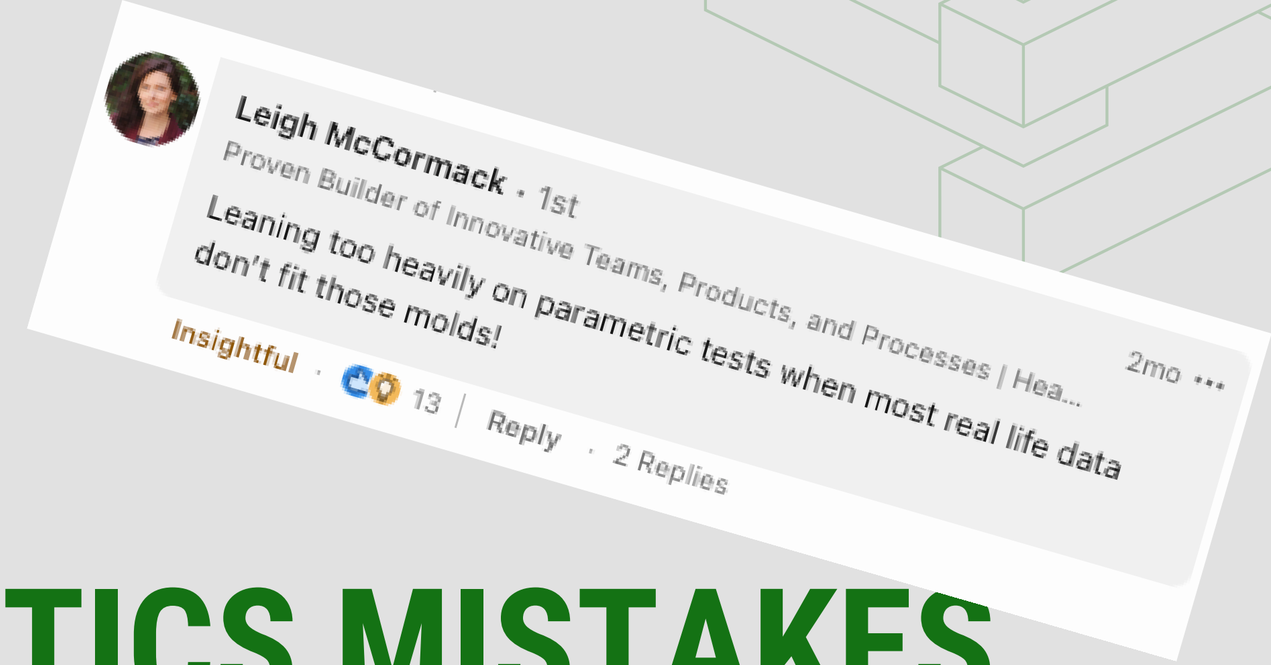
THE MOST UNDER-APPRECIATED ASPECT OF DATA SCIENCE

BIGGEST STATISTICS MISTAKES IN DATA SCIENCE



LST STATISTICS MISTAKES IN DATA SCIENCE





10 STATISTICS MISTAKES IN DATA SCIENCE

10 STATISTICS MISTAKES IN DATA SCIENCE

Jessie Lamontagne • 1st
Machine Learning | Data Science | Analytics
2mo (edited)

Describing only the most vanilla undergrad-level regression as Regression, looking to graph neural networks as the only way to get causal inference while ignoring decades of research on causal inference with tabular data, treating point estimates as truth, not calculating uncertainty at all "because big data".

Celebrate · 23 | Reply · 10 Replies

Leigh McCormack • 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Rodrigo Rivera • 1st
Data Products for Business Users Builder | Enabling Anyone t...
2mo

Basing all decisions around the p-value

Like · 9 | Reply · 3 Replies

10 STATISTICS MISTAKES IN DATA SCIENCE

Jessie Lamontagne • 1st
Machine Learning | Data Science | Analytics
2mo (edited)

Describing only the most vanilla undergrad-level regression as Regression, looking to graph neural networks as the only way to get causal inference while ignoring decades of research on causal inference with tabular data, treating point estimates as truth, not calculating uncertainty at all "because big data".

Celebrate · 23 | Reply · 10 Replies

Leigh McCormack • 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Kristen Kehrer (She/Her) • 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo

Not so much fellow statisticians, but I was a broken record saying "correlation is not causation we need to test these results" to stakeholders after sharing model output.

Like · 9 | Reply · 1 Reply

Rodrigo Rivera • 1st
Data Products for Business Users Builder | Enabling Anyone t...
2mo

Basing all decisions around the p-value

Like · 9 | Reply · 3 Replies

LST STATISTICS M
IN DATA SCIEN

Jessie Lamontagne • 1st
Machine Learning | Data Science | Analytics
2mo (edited)

Describing only the most vanilla undergrad-level regression as Regression, looking to graph neural networks as the only way to get causal inference while ignoring decades of research on causal inference with tabular data, treating point estimates as truth, not calculating uncertainty at all "because big data".

Celebrate · 23 | Reply · 10 Replies

Leigh McCormack • 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Peter Mancini (He/Him) • 1st
Principal at Stealth Startup
2mo

1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.

3. No concept of uncertainty. I think this is a failure that goes through a much wider swathe of practitioners including fairly seasoned practitioners. In fact, all of the faults that you mentioned contain uncertainty as a key feature. You were going to be uncertain about every single outcome, sometimes a lot, sometimes a miniscule amount, usually something in between. Knowing it will help.

Celebrate · 3 | Reply · 1 Reply

Kristen Kehrer (She/Her) • 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo

Not so much fellow statisticians, but I was a broken record saying "correlation is not causation we need to test these results" to stakeholders after sharing model output.

Like · 9 | Reply · 1 Reply

Rodrigo Rivera • 1st
Data Products for Business Users Builder | Enabling Anyone t...
mo

Basing all decisions around the p-value

Like · 9 | Reply · 3 Replies

LST STATISTICS M
IN DATA SCIEN

Jessie Lamontagne • 1st
Machine Learning | Data Science | Analytics
2mo (edited)

Describing only the most vanilla undergrad-level regression as Regression, looking to graph neural networks as the only way to get causal inference while ignoring decades of research on causal inference with tabular data, treating point estimates as truth, not calculating uncertainty at all "because big data".

Celebrate · 23 | Reply · 10 Replies

Leigh McCormack • 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Peter Mancini (He/Him) • 1st
Principal at Stealth Startup
2mo

1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.

3. No concept of uncertainty. I think this is a failure that goes through a much wider swathe of practitioners including fairly seasoned practitioners. In fact, all of the faults that you mentioned contain uncertainty as a key feature. You were going to be uncertain about every single outcome, sometimes a lot, sometimes a miniscule amount, usually something in between. Knowing it will help.

Celebrate · 3 | Reply · 1 Reply

Kristen Kehrer (She/Her) • 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo

Not so much fellow stat... but I was a broken record saying "correlation is not caus... results" to stakeholders after shar...

Like · 9 | Reply

Rodrigo Rivera • 1st
Data Products for Business Users Builder | Enabling Anyone t...
mo

...s around the p-value

...ly · 3 Replies

Joe Wyer (He/Him) • 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo

Thinking statistical significance should be the filter for go/no go of business decisions.

Like · 4 | Reply · 2 Replies

LST STATISTICS M
IN DATA SCIEN

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo
Using SMOTE when they should not 😂
Celebrate · 2 | Reply · 1 Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Describing only the most vanilla undergrad-level regre
Regression, looking to graph neural networks as the only way to s
causal inference while ignoring decades of research on causal
inference with tabular data, treating point estimates as truth, not
calculating uncertainty at all "because big data".
Celebrate · 23 | Reply · 10 Replies

Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo
Leaning too heavily on parametric tests when most real life data
don't fit those molds!
Insightful · 13 | Reply · 2 Replies

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
2mo
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.

3. No concept of uncertainty. I think this is a failure that goes through a much wider swathe of practitioners including fairly seasoned practitioners. In fact, all of the faults that you mentioned contain uncertainty as a key feature. You were going to be uncertain about every single outcome, sometimes a lot, sometimes a miniscule amount, usually something in between. Knowing it will help.
Celebrate · 3 | Reply · 1 Reply

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo
Not so much fellow stat
but I was a broken record saying
"correlation is not caus
results" to
stakeholders after shar
Like · 9 | Reply

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo
Thinking statistical significance should be the filter for go/no go of business decisions.
Like · 4 | Reply · 2 Replies

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
mo
around the p-value
3 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo

Using SMOTE when they should not 😂

Celebrate · 2 | Reply · 1 Reply

Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
2mo

1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.

3. No concept of uncertainty. I think this is a failure that goes through a much wider swathe of practitioners including fairly seasoned practitioners. In fact, all of the faults that you mentioned contain uncertainty as a key feature. You were going to be uncertain about every single outcome, sometimes a lot, sometimes a miniscule amount, usually something in between. Knowing it will help.

Celebrate · 3 | Reply · 1 Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics

Describing only the most vanilla undergrad-level regre... Regression, looking to graph neural networks as the only way to s... causal inference while ignoring decades of research on causal inference with tabular data, treating point estimates as truth, not calculating uncertainty at all "because big data".

Celebrate · 23 | Reply · 10 Replies

Hai Tran (He/Him) · 1st
Data Science and Analytics Specialist
2mo

Successfully rejecting or failing to reject Null Hypothesis H0 does not prove that the Alternative Hypothesis is True or False. We can only say whether H1 is supported by the observed data or not.

Another common pitfall should be being mistaken for the correlation between 2 variables and thinking that it is a causation relationship between these.

Celebrate · 4 | Reply · 1 Reply

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo

Not so much fellow stat... but I was a broken record saying "correlation is not caus... results" to stakeholders after shar...

Like · 9 | Reply

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
mo

around the p-value

ly · 3 Replies

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo

Thinking statistical significance should be the filter for go/no go of business decisions.

Like · 4 | Reply · 2 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo

Using SMOTE when they should not 😂

Celebrate · 2 | Reply · 1 Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics

Describing only the most vanilla undergrad-level regre... looking to graph neural networks as the only way to ... decades of research on causal ...ates as truth, not ...

Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Aaron Sheldon · Following
VP of Engineering at Genstate
2mo (edited)

In general across all scientific fields it is failing to understand the strong limitations of the mathematical assumptions of statistical models. Basically applying statistical testing as a form of exploratory analysis. Or more bluntly assuming regression and its generalizations all the way up to deep learning are a magical oracle that can tell you how your system works.

Statistical tests and models assume a priori that you have a theory of how the system your are studying works. They simply estimate the quantitative bounds of your theory. Not generally whether it is right or wrong. This is even true of classical hypothesis testing.

Celebrate · 34 | Reply · 4 Replies

Hai Tran (He/Him) · 1st
Data Science and Analytics Specialist
2mo

Successfully rejecting or failing to reject Null Hypothesis H0 does not prove that the Alternative Hypothesis is True or False. We can only say whether H1 is supported by the observed data or not.

Another common pitfall should be being mistaken for the correlation between 2 variables and thinking that it is a causation relationship between these.

Celebrate · 4 | Reply · 1 Reply

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
2mo

1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

2. Relying on normal distribution tools de facto. Better to rely on tools that will work with the distribution of the data that you have.

3. No concept of uncertainty. I think this is a failure that goes through a much wider swathe of practitioners including fairly reasoned practitioners. In fact, all of the faults that you mentioned contain uncertainty as a key feature. You were going to be uncertain about every single outcome, sometimes a lot, sometimes a miniscule amount, usually something in between. Knowing it will help.

Celebrate · 3 | Reply · 1 Reply

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo

Not so much fellow stat... but I was a broken record saying "correlation is not caus... results" to stakeholders after shar...

Like · 9 | Reply

Rodrigo Rivera · 1st
🦆 Data Products for Business Users Builder | 🦊 Enabling Anyone t...
mo

Racing ... around the p-value

...ly · 3 Replies

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo

Thinking statistical significance should be the filter for go/no go of business decisions.

Like · 4 | Reply · 2 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo

Using SMOTE when they should not 😂

Celebrate · 2 | Reply · 1 Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics

Describing only the most vanilla undergrad-level regre... looking to graph neural networks as the only way to ... decades of research on causal ...nates as truth, not ...

Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo

Leaning too heavily on parametric tests when most real life data don't fit those molds!

Insightful · 13 | Reply · 2 Replies

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
2mo

1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

Aaron Sheldon · Following
VP of Engineering at Genstate
2mo (edited)

In general across all scientific fields it is failing to understand the strong limitations of the mathematical assumptions of statistical models. Basically applying statistical testing as a form of exploratory analysis. Or more bluntly assuming regression and its generalizations all the way up to deep learning are a magical oracle that can tell you how your system works.

Statistical tests and models assume a priori that you have a theory of how the system your are studying works. They simply estimate the quantitative bounds of your theory. Not generally whether it is right or wrong. This is even true of classical hypothesis testing.

Celebrate · 34 | Reply · 4 Replies

Hai Tran (He/Him) · 1st
Data Science and Analytics Specialist
2mo

Successfully rejecting or failing to reject Null Hypothesis H0 does not prove that the Alternative Hypothesis is True or False. We can only say whether H1 is supported by the observed data or no...

Another common pitfall should be being mistaken for the ... correlation between 2 variables and thinking that it is a c... relationship between these.

Celebrate · 4 | Reply · 1 Reply

Tyler Buffington, PhD · 1st
Senior Data Scientist
2mo

1. Not understanding decision analysis and value of information (VOI). For example, building an extremely complicated model when a simple analysis is sufficient to guide the business to the correct decision.

2. Failing to understand that underpowered experiments are untrustworthy even if the result is statistically significant.

Insightful · 6 | Reply · 2 Replies

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate @ CometML | The Cool Data P...
2mo

Not so much fellow stat... but I was a broken record saying "correlation is not caus... ...results" to stakeholders after shar...

Like · 9 | Reply

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...

...around the p-value

...ly · 3 Replies

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo

Thinking statistical significance should be the filter for go/no go of business decisions.

Like · 4 | Reply · 2 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo
Using SMOTE when they should not 😂
Celebrate · 2 | Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Describing only the most vanill...

Thomas Speidel · 2nd
Statistician and Data Scientist
2mo
1. The independent/predictor variables have to be normally distributed
2. Linear regression can only fit a straight line
3. Logistic regression is a classifier
4. One should always split the data into training and test sets and
5. Cross-validation is the only resampling method
6. Outliers need to be removed
7. One should always remove non-significant predictors
8. R^2 is the only performance metric to look at
9. Doing a vanilla least square regression because that's what we remember from stat 101, and then jumping to a NN with 100 hidden layers and 12 hours of training time because the linear regression didn't perform well
10. That explaining a model simply means doing a feature importance/SHAP plot
12. Not thinking causally
13. Not quantifying uncertainty of the model and predictions
14. Not doing literature reviews (remembering that data science is a vast field that grew in parallel but independent fields)
15. How unstable and sometimes problematic feature selection algos are (at least for weak features)
Like · 13 | Reply · 2 Replies

Aaron Sheldon · Following
VP of Engineering at Genstate
In general across all scienti
strong limitations of the mat
models. Basically applying st
exploratory analysis. Or more
generalizations all the way up t
that can tell you how your syste
Statistical tests and models assun
of how the system your are studyin
the quantitative bounds of your theo
right or wrong. This is even true of cl
Celebrate · 34 | Reply · 4 Replie

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate
Not so much fellow stat
"correlation is not caus
stakeholders after shar
Like · 9 | Reply

Specialist
2mo
failing to reject Null Hypothesis H0 does
tive Hypothesis is True or False. We can
ported by the observed data or no
d be being mistaken for the
s and thinking that it is a c
Reply

Leigh McCormack · 1st
Proven Builder of Innovative Teams, Products, and Processes | Hea...
2mo
Leaning too heavily on parametric tests when most real life data don't fit those molds!
Insightful · 13 | Reply · 2 Replies

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
2mo
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

Tyler Buffington, PhD · 1st
Senior Data Scientist
2mo
1. Not understanding decision analysis and value of information (VOI). For example, building an extremely complicated model when a simple analysis is sufficient to guide the business to the correct decision.
2. Failing to understand that underpowered experiments are untrustworthy even if the result is statistically significant.
Insightful · 6 | Reply · 2 Replies

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
...s around the p-value
...ly · 3 Replies

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo
Thinking statistical significance should be the filter for go/no go of business decisions.
Like · 4 | Reply · 2 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo
Using SMOTE when they should not 😂
Celebrate · 2 | Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Describing only the most vanill...

Thomas Speidel · 2nd
Statistician and Data Scientist
2mo
1. The independent/predictor variables have to be normally distributed
2. Linear regression can only fit a straight line
3. Logistic regression is a classifier
4. One should always split the data into training and test sets and
5. Cross-validation is the only resampling method
6. Outliers need to be removed
7. One should always remove non-significant predictors
8. R^2 is the only performance metric to look at
9. Doing a vanilla least square regression because that's what we remember from stat 101, and then jumping to a NN with 100 hidden layers and 12 hours of training time because the linear regression didn't perform well
10. That explaining a model simply means doing a feature importance/SHAP plot
12. Not thinking causally
13. Not quantifying uncertainty of the model and predictions
14. Not doing literature reviews (remembering that data science is a vast field that grew in parallel but independent fields)
15. How unstable and sometimes problematic feature selection algos are (at least for weak features)
Like · 13 | Reply · 2 Replies

Aaron Sheldon · Following
VP of Engineering at Genstate
In general across all scienti
strong limitations of the mat
models. Basically applying st
exploratory analysis. Or more
generalizations all the way up t
that can tell you how your syste
Statistical tests and models assun
of how the system your are studyin
the quantitative bounds of your the
right or wrong. This is even true of cl
Celebrate · 34 | Reply · 4 Replie

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate
Not so much fellow stat
"correlation is not caus
stakeholders after shar
Like · 9 | Reply

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo
Thinking statistical significance should be the filter for go/no go of business decisions.
Like · 4 | Reply · 2 Replies

Leigh McCorm...
Proven ...
Insig...

Jeff Bean · 2nd
Senior Consultant at Deloitte
2mo
Failing to recognize the difference between "significant" (i.e., P<0.05, not likely due to chance, etc.) and "meaningful" (i.e., Cohen's delta, worthwhile, cost-effective, etc.). You need both.
Celebrate · 2 | Reply

...Specialist
failing to reject Null Hypothesis H0 does
tive Hypothesis is True or False. We can
ported by the observed data or n
d be being mistaken for the
s and thinking that it is a c
...ply

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

Tyler Buffington, PhD · 1st
Senior Data Scientist
2mo
1. Not understanding decision analysis and value of information (VOI). For example, building an extremely complicated model when a simple analysis is sufficient to guide the business to the correct decision.
2. Failing to understand that underpowered experiments are untrustworthy even if the result is statistically significant.
Insightful · 6 | Reply · 2 Replies

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
...around the p-value
...ly · 3 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo
Using SMOTE when they should not 😂
Celebrate · 2 | Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Describing only the most vanilla ...

Thomas Speidel · 2nd
Statistician and Data Scientist
1. The independent/predictor variables have to be normally distributed
... linear regression can only fit a straight line
... test sets and

Aaron Sheldon · Following
VP of Engineering at Genstate
In general across all scienti
strong limitatio...

Justyn Hornor (He/Him) · 2nd
Fractional CTO/CPO. Polymath. Inventor. Technology generalist. Str...
2mo
Not recognizing Type 1 vs Type 2 errors as it relates to the org's needs.
Like · 3 | Reply
...ve bounds of your the
right or wrong. This is even true of cl
Celebrate · 34 | Reply · 4 Replie

...'s what we
...h 100 hidden
... regression
...ure
didn't perform well
10. That explaining a model simply means doing a ...
importance/SHAP plot
12. Not thinking causally
13. Not quantifying uncertainty of the model and predictions
14. Not doing literature reviews (remembering that data science is a vast field that grew in parallel but independent fields)
15. How unstable and sometimes problematic feature selection algos are (at least for weak features)
Like · 13 | Reply · 2 Replies

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate
Not so much fellow stat
"correlation is not caus
stakeholders after shar
Like · 9 | Reply

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo
Thinking statistical significance should be the filter for go/no go of business decisions.
Like · 4 | Reply · 2 Replies

Leigh McCorm...
Proven ...
L...
do...
Insig...

Jeff Bean · 2nd
Senior Consultant at Deloitte
2mo
Failing to recognize the difference between "significant" (i.e., P<0.05, not likely due to chance, etc.) and "meaningful" (i.e., Cohen's delta, worthwhile, cost-effective, etc.). You need both.
Celebrate · 2 | Reply
...eplies

...r life data

...t
...s Specialist
...failing to reject Null Hypothesis H0 does
...tive Hypothesis is True or False. We can
...ported by the observed data or n...
...d be being mistaken for the
...s and thinking that it is a c...
...ply

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.

Tyler Buffington, PhD · 1st
Senior Data Scientist
2mo
1. Not understanding decision analysis and value of information (VOI). For example, building an extremely complicated model when a simple analysis is sufficient to guide the business to the correct decision.
2. Failing to understand that underpowered experiments are untrustworthy even if the result is statistically significant.
Insightful · 6 | Reply · 2 Replies

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
...s around the p-value
...ly · 3 Replies

Zion Pibowei · 1st
Principal Data Scientist | Head of Data Science @Periculum | Data ...
2mo
Using SMOTE when they should not 😂
Celebrate · 2 | Reply

Jessie Lamontagne · 1st
Machine Learning | Data Science | Analytics
Describing only the most vanilla

Thomas Speidel · 2nd
Statistician and Data Scientist
1. The independent/predictor variables have to be normally distributed

Aaron Sheldon · Following
VP of Engineering at Genstate
In general across all scienti
strong limitatio

Justyn Hornor (He/Him) · 2nd
Fractional CTO/CPO. Polymath. Inventor. Technolo
Not recognizing Type 1 vs Type 2 errors as needs.
Like · 3 | Reply
right or wrong. This is even true of cl
Celebrate · 34 | Reply · 4 Replie

Kristen Kehrer (She/Her) · 1st
Data Scientist | Developer Advocate
Not so much fellow stat
"correlation is not caus
stakeholders after shar
Like · 9 | Reply

Timothy Dobbins · You
Building data products @ Gridsearch | Principal DS @ Trilliant Health
2mo · Edited

Statistics is the biggest skills gap in data science right now. I thought it was software engineering but having seen how data science is done in many organizations and hearing from statisticians yesterday about the common mistakes data scientists make, I'm convinced that the coding gaps are nothing compared to the stats gaps.

My stats friends really showed up yesterday.

Mary van Valkenburg and 414 others
63 comments · 31 reposts

Like · 13 | Reply · 2 Replies

Joe Wyer (He/Him) · 1st
Head of Economics and Applied Science @ Haus | We're Hiring!
1mo
Thinking statistical significance should be the filter for go/no go of business decisions.
Like · 4 | Reply · 2 Replies

Leigh McCorm
Proven
Insig

Jeff Bean · 2nd
Senior Consultant at Deloitte
2mo
Failing to recognize the difference between "significant" (i.e., P<0.05, not likely due to chance, etc.) and "meaningful" (i.e., Cohen's delta, worthwhile, cost-effective, etc.). You need both.
2 | Reply

Peter Mancini (He/Him) · 1st
Principal at Stealth Startup
1. Not knowing what distribution their data is and using tools that depend upon a specific distribution, usually the normal distribution.
2mo

uffington, PhD · 1st
ta Scientist
derstanding decision analysis and value of information example, building an extremely complicated model when alysis is sufficient to guide the business to the correct
2mo
nderstand that underpowered experiments are even if the result is statistically significant.
Insightful · 6 | Reply · 2 Replies

Rodrigo Rivera · 1st
Data Products for Business Users Builder | Enabling Anyone t...
around the p-value
3 Replies

REQUIREMENTS ARE HARD TO DEFINE BECAUSE OF **AMBIGUITY**

SUCCESS IS HARD TO DEFINE BECAUSE OF **UNCERTAINTY**

VALUE IS MIS-MEASURED BECAUSE OF **BAD SCIENCE**

AMBIGUITY OCCURS WHEN WE DON’T KNOW WHAT QUESTIONS TO ASK

UNCERTAINTY OCCURS WHEN WE DON’T KNOW HOW TO VALIDATE THE ANSWERS

BAD SCIENCE OCCURS WHEN WE INCORRECTLY VALIDATE THE ANSWERS.

THANK YOU!

LET'S CONNECT