Math and statistics and bears: oh my

Timothy Petersen
Math and Statistics

MATH AND STATISTICS AND BEARS

Tim Petersen, PhD

OH MY!

WHY ARE WE HERE?

- "Math is hard"
- Don't be Barbie.

LET'S START SIMPLE

- A linear equation
  \( y = mx + b \)

A LITTLE MORE...

- Quadratic form
  \( y = ax^2 + bx + c \)
We’ve seen an exponent already ($2^3$)

Logarithms are the same thing, but in reverse

If $b^y = x$, then $y = \log_b(x)$ So the question is: to what power must we raise $b$ to get $x$?

- Example: $1000 = 10^3 = 10 \times 10 \times 10$, so $\log_{10}(1000) = 3$

- pH is a logarithmic scale, so was the Richter scale
THE MOST COMMON BASES

- Any number can be a base, but two are most common
  - Base 10 (aka “common logarithm”)
    - \( \log(x) \)
  - Base \( e \) (aka “natural logarithm”)
    - \( \ln(x) \)
    - \( e \) is interesting

A LOGARITHMIC GRAPH

- \( y = \log_2(x) \)
- Hmmmm- that shape looks familiar...

YES IT DOES

- Logarithms and exponents are inverse operations, so their graphs are inverses around \( y = x \).
THE NORMAL DISTRIBUTION

\[ y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

ITS PERCENTAGES

- 99.73% of the data falls within 3 standard deviations of the mean.
- 68.27% of the data falls within 1 standard deviation of the mean.
- 95.45% of the data falls within 2 standard deviations of the mean.

![Normal Distribution Graph](image)
LOGNORMAL

- Sometimes X isn't normally distributed, but \( \log(x) \) is
- Drug elimination is a logarithmic function

CUMULATIVE DISTRIBUTION FUNCTION

SURVIVAL GRAPH

- Complement of cumulative distribution function
- \( 1 - \text{CDF} \)

INTRO TO STATS

In search of a tiny \( p \)
OKAY, LET’S START EASY

- **Probability**
  - The study of chance and likelihood
  - A probability is a number expressing a likelihood: \(0 \leq P \leq 1\)

- Probability: Given the information in the pail, what is in your hand?

- Statistics: Given the information in your hand, what is in the pail?

**CONDITIONAL PROBABILITY**

- If we flip a coin twice, what is the probability of **2 heads**?
  * A = heads on first toss; B = heads on second toss
  * \(P(A \text{ and } B): P(A \cap B) = P(A) \times P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}\)

- What is the probability of rolling a **1 or 2** on a 6-sided die?
  * \(P(A \text{ or } B)\) if mutually exclusive: \(P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}\)

- What is the probability of drawing a **queen** or a **heart** card?
  * \(P(A \text{ or } B)\) if not exclusive: \(P(A \cup B) = P(A) + P(B) − P(A \cap B) = \frac{4}{52} + \frac{13}{52} − \frac{1}{52} = \frac{16}{52} = \frac{4}{13}\)

**SENSITIVITY / SPECIFICITY**

- These familiar concepts are probabilistic statements

- Because of that, we can treat them as a “probability tree” of conditional probabilities.

- A phenomenon you’ve heard about...
A REASONABLY ACCURATE TEST...

- Sensitivity 95%
- Specificity 90%

POPULATION AND SAMPLE

POPULATION

- The group of interest
- We want to make supported conclusions about them as a group
  - All humans
  - Women in labor
  - Neurosurgery patients with no comorbidities
  - Adult women with T1DM
- Can’t measure whole groups. With stats, we don’t need to

THE SAMPLE

- The people used in this study
  - For valid conclusions, the sample must match the population well
  - Single biggest factor: random sampling
UNIVARIATE STATISTICS

Just describe

MEASURES OF CENTRAL TENDENCY

- Mode: Most common value
- Median: Middle observation (or average of middle 2); 50th percentile.
  - Interquartile range: 25th and 75th percentiles
- "Average" (arithmetic mean): sum of observations, divided by count of observations
  \[
  \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
  \]
How spread out are the observations?

\[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \]

Variance = \( s^2 \)

Variance = \( s^2 \)

How far is this particular sample mean likely to be from the true population mean?

- Our sample should more or less mimic the population, but it’s likely to be a little off. With a larger sample, we should be less “off.”
- Reasonable to think of it as a “standard deviation for our estimate of the population mean.”

\[ SE_{\bar{x}} = \frac{s}{\sqrt{n}} \]

Our estimate of some population parameter (e.g. \( \mu \), the population mean) probably isn’t exactly right, but gives us some ballpark.

- Larger SE: wider range of possibilities
- Smaller sample: wider range of possibilities

- Stated with % confidence level, e.g. 95%
  - Example: observed mean 12, 95% CI: (7.8, 16.2)
  - If we took many samples of this size, 95% of the resulting confidence intervals should contain the true population mean

For a 95% CI of the mean, \( \bar{x} \pm 1.96SE_{\bar{x}} \)

99%: \( \bar{x} \pm 2.58SE_{\bar{x}} \)
TYPES OF DATA (OLD SCHOOL)

- Categorical
  - Nominal
    - Not ordered
    - Male vs. female
  - Ordinal...
    - Ordered, but uninterpretable arithmetic
    - Good, fair, poor
    - Pain scores

- Continuous
  - Interval
    - +/- differences are consistent, but zero isn’t meaningful, so ratios aren’t
    - Most temp scales
  - Ratio
    - Consistent differences, meaningful zero and ratios
    - Linear measurements, °K, body mass, solution concentration, ...

THAT WAS THEN

- When possible, more precise measurements are usually better
- Modern usage worries less about the label than about appropriate statistical strategies.
  - Ordinal data of good, fair, poor might be best handled with analyses usually used for categorical data
  - Pain scores can sometimes be handled with some continuous-data analyses

WHAT THE P-VALUE?

HYPOTHESIS TESTING
If there really isn’t a difference between groups in the population, how often would a sample of this size show a difference this big or bigger?

A small p value lets you infer that the apparent difference between groups is probably not due to random chance alone.

But note: statistical significance ≠ clinical significance

Science does not prove statements to be true. “Proof” is a matter for mathematicians, lawyers, and other drunkards.

Science proceeds by rejecting the null hypothesis: usually that’s “no difference.”

“If there really is no difference between populations, I’d expect to see at least this strong (p x 100)% of the time. That’s so unlikely that I can provisionally reject that idea, and behave as if there really is a difference.”

The traditional threshold for statistical significance (α) is 0.05. This corresponds to a 5% risk of false positives.

Risky treatments: often use p = 0.01
Lower stakes: why not use p = 0.10

Must decide before conducting the study.
Of one state’s 1041 mesothelioma patients, 1024 report exposure to asbestos.

The state sampled 17983 non-MT patients: 3013 of them reported asbestos exposure.

How should we proceed?

Is there an association between asbestos exposure status and mesothelioma?

Uses count data (categorical)

A bit of a blunt instrument: does not quantify strength of association; just significance for “departure from random”

<table>
<thead>
<tr>
<th></th>
<th>Dx mesothelioma</th>
<th>No diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed to asbestos</td>
<td>1024</td>
<td>3013</td>
</tr>
<tr>
<td>Not exposed to asbestos</td>
<td>17</td>
<td>14970</td>
</tr>
</tbody>
</table>

Chi-square compares the observed counts to those that would be expected if there were no association.

Any cells with expected count = 0

If 2x2 table, any cells with expected cell count <5 (some say 10)

Larger table: >20% of cells with expected cell count <5
### WHAT THEN?

- **Fisher’s Exact**
  - Highly computation-intensive
  - The computer literally finds all possible tables with the same number of observations
  - What percentage of them have a result more extreme than the one observed?
  - That’s the p value

### CAN WE DO BETTER?

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>No disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Not exposed</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

- **Relative Risk**
  - Usually for **prospective** studies
  - E.g. “Exposure X triples the risk of disease Y”

\[
\frac{a}{a + b} \quad \frac{c}{c + d}
\]

### YES WE CAN

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>No disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Not exposed</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

- **Odds Ratio**
  - Best for retrospective (case-control) studies, like our MT example
  - We don’t know how common the disease is, or the exposure rate; that’s OK

\[
\frac{ad}{bc}
\]

### WHICH TO USE?

- **RR for prospective studies** (randomized treatment groups)
- **OR for case-control studies** (take ’em as they come)
- OR underlies other areas of stats, such as logistic regression
- OR relationships are invertible, unlike RR
- RR is more intuitive, but also abusable (very rare conditions)
**ANOTHER SITUATION**

- A high school track and field team averaged 15 meters on shot put.
- Their coaches averaged 17 meters.
- Are the coaches better?

---

**T-TEST**

- Do these two samples have similar means?
- Is 37°C really the population’s mean body temperature?

**Assumptions:**
- Data are normally distributed in both samples
- The two samples have the same variance and size
  - There are versions of t-test that account for violations of those

**What does it do?**
- Compares difference between groups to the total variation within the pooled groups

**PAIRED T-TEST**

- When there are pre- and post-tests on the same subjects, we have a new layer of information: the scores are **paired**.
  - After being beaten by the coaches, the track team had extra practice. Their average after that practice was 18.

- Paired t-test: take the paired scores’ differences, and compare them against the null hypothesis of difference=0.

- When possible, it’s much more powerful than a regular t-test.
I want to do a t-test, but my data aren’t normal!

WMW test has a clever workaround:
1. Convert pooled data to rank values
2. Do a regular t-test on the ranks
3. Profit!

Tests like this are called *nonparametric*, because they don’t rely on the assumption of a particular parameter, e.g. a normal distribution.

On average, anesthesiologists at Peds OR, Main OR, and OSIS spend 15, 20, and 13 minutes with patients (respectively) in pre-op.

Is this indicative of a real difference in practice?

If we have \(k > 2\) groups, we can just do t-tests on all the pairs, right? OSIS v. Peds, Main v. Peds, OSIS v. Main

For \(k\) groups, the number of pairs is \(\frac{k!}{2(k-2)!} = \frac{k(k-1)}{2}\)

5 groups: 10 pairs
7 groups: 21 pairs
10 groups: 45 pairs

Tedious. And....

Conditional probabilities!
If we’re using a \(p < 0.05\) threshold once, we have a 95% chance of being “right.”

If we’re doing test after test after test, being “right” on all of them comes out to \(0.95 \times 0.95 \times 0.95\)....

At 10 tests (5 groups), we have only a 60% chance.
At 21 tests (7 groups), it’s 34%.
At 45 tests, (10 groups), 10%: that’s a 90% chance of at least 1 spurious result.
**ANOVA**

- **Analysis of Variance**
- Very powerful, versatile test
- Generalization of t-test to >2 groups, preserves low error risk
  - Same assumptions: normality, similar variances and sample sizes

---

**NONPARAMETRIC ANOVA**

- Kruskal-Wallis
  - Same idea as WMW: do it on the ranks

---

**AND NOW FOR SOMETHING COMPLETELY DIFFERENT**

- I have data on stature and long bone lengths from US servicemen. Can I use these data to predict stature for John/Jane Doe cases with skeletonized remains?
- Is there a relationship between BMI and length of hospital stay for joint-replacement patients?

---

**CORRELATION AND REGRESSION**

- I have 2 continuous variables. Are they related?
  - Correlation: just the strength of relationship
  - Regression: that, plus the equation for the “regression line,” usually for using X to predict Y
- Big question: is the slope of the regression line = 0?
**IT AIN’T CAUSATION**

- Things that correlate:
  - Internet traffic and autism diagnoses
  - Shoe size and reading ability
  - Sleeping with shoes on and awakening with a hangover
  - Ice cream sales and drowning risk
  - Atmospheric CO₂ and obesity
  - Cancer and ________
  - Pirates and global warming (negative correlation)

**ALWAYS LOOK AT THE DATA**

- The next slide has 4 data distributions

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of ( x )</td>
<td>9 (exact)</td>
</tr>
<tr>
<td>Variance of ( x )</td>
<td>11 (exact)</td>
</tr>
<tr>
<td>Mean of ( y )</td>
<td>7.50 (to 2 decimal places)</td>
</tr>
<tr>
<td>Variance of ( y )</td>
<td>4.122 or 4.127 (to 3 decimal places)</td>
</tr>
<tr>
<td>Correlation between ( x ) and ( y )</td>
<td>0.816 (to 3 decimal places)</td>
</tr>
<tr>
<td>Linear regression line</td>
<td>( y = 3.00 + 0.500x ) (to 2 and 3 decimal places, respectively)</td>
</tr>
</tbody>
</table>

**Global Average Temperature Vs. Number of Pirates**

![Graph showing correlation between global average temperature and number of pirates](venganza.org)

**Linear regression line**

- \( y = 3.00 + 0.500x \) (to 2 and 3 decimal places, respectively)
BUT WHAT IF...

- What if my data aren’t appropriate?
- Instead of traditional Pearson correlation (r), use Spearman’s ($\rho$)
  - That ranks trick again
  - Effectively a test for “monotonicity”

OTHER VERSIONS

- Logistic regression:
  - X-variable is continuous, Y-variable is binary (odds)
- Analysis of Covariance (ANCOVA):
  - Is the linear relationship between X- and Y-variables different between ≥2 groups?

AFTERWARD

- Statistical tests are only valid if they are specified a priori.
  - As in: “We will gather data, and then test these specific hypotheses with these particular analyses.”
- The p value for most post-hoc tests cannot be interpreted.
  - We don’t know how many post-hoc tests were done
  - Scanning the data for a promising trend, and then analyzing it = running a very large number of analyses and picking the “best”

POWER ANALYSIS

More power, Scotty!
I CAN'T GET NO SATISFACTION

- A small sample might not adequately represent the population.
- A huge one almost certainly does.
- At what point are we satisfied that we have enough?

TYPES OF ERROR

- Type I
  - Wrongly "seeing" an effect that isn't really there (false positive)
  - Risk = \(\alpha\)
  - If your \(p\) value is < \(\alpha\), you win: statistical significance

- Type II
  - Wrongly "missing" a real effect (false negative)
  - Risk = \(\beta\)
  - \(1 - \beta\) = power
  - Power is the chance of "seeing" an effect if there is one

ANOTHER LOOK

<table>
<thead>
<tr>
<th></th>
<th>Treatment is Effective</th>
<th>Treatment is Ineffective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive study</td>
<td>True Positive</td>
<td>False Positive; (T1; \alpha)</td>
</tr>
<tr>
<td>Negative study</td>
<td>False Negative; (T2; \beta)</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

MOST COMMON USE OF POWER

- Sample size estimation
- "How many patients do we need to have in this study?"
- Calculation involves:
  - Size of effect (difference between groups) and
  - Background variation (variation within groups)
- Many ways to get these figures
  - Pilot study
  - Literature review
  - Estimate of minimum clinically-significant difference
I want 80% power for my study, and will use the usual 0.05 significance threshold.
Comparing 2 means, so I’ll use t-test
A similar published study says that $s = 2.5$, and the data were normally distributed (yay!)
They found a difference between the means of 4.
With these parameters, I need 16 subjects (8/group).
But the clinically significant difference is more like 2.
I’d still be interested in an effect that small. But my original sample is underpowered to detect it! (has 32% power)
Bigger sample: 52 gives me 80% power for this smaller effect.

Never
- do a power analysis after a negative study, in an effort to “prove” no difference (“post-hoc power”)
- This is meaningless.

What you can do
- Use data from a negative study to set the sample size for a future study
---

**LET’S INVITE EVERYBODY**

- Studies often disagree
- So let’s stack ‘em together

<table>
<thead>
<tr>
<th>Study</th>
<th>Odds ratio [95% CI]</th>
<th>% Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aikensland</td>
<td>1.23 (0.89, 1.68)</td>
<td>5.2</td>
</tr>
<tr>
<td>Feiner</td>
<td>0.20 (0.14, 0.29)</td>
<td>6.4</td>
</tr>
<tr>
<td>Skopak</td>
<td>0.27 (0.23, 0.33)</td>
<td>6.9</td>
</tr>
<tr>
<td>Durrer</td>
<td>0.11 (0.02, 0.36)</td>
<td>6.9</td>
</tr>
<tr>
<td>Ebbel</td>
<td>1.27 (0.43, 3.47)</td>
<td>10.1</td>
</tr>
<tr>
<td>Fejer</td>
<td>1.15 (0.42, 4.23)</td>
<td>2.5</td>
</tr>
<tr>
<td>Gocher</td>
<td>1.37 (0.44, 4.30)</td>
<td>9.2</td>
</tr>
<tr>
<td>Kollerberg</td>
<td>1.20 (0.77, 2.17)</td>
<td>6.7</td>
</tr>
<tr>
<td>Kay</td>
<td>0.26 (0.17, 0.29)</td>
<td>6.1</td>
</tr>
<tr>
<td>Klocke</td>
<td>0.49 (0.36, 0.72)</td>
<td>5.6</td>
</tr>
<tr>
<td>Vencel</td>
<td>0.11 (0.07, 0.21)</td>
<td>4.2</td>
</tr>
<tr>
<td>Odersky</td>
<td>1.28 (0.58, 2.87)</td>
<td>6.7</td>
</tr>
<tr>
<td>Stina</td>
<td>0.17 (0.10, 0.29)</td>
<td>6.9</td>
</tr>
<tr>
<td>Valiron</td>
<td>1.42 (0.72, 2.80)</td>
<td>7.3</td>
</tr>
</tbody>
</table>

---

**WAIT... UNPUBLISHED?**

- Why does it work?
  - Effectively a larger sample
  - Handles variation between studies
  - Synthesis

- Why doesn’t it?
  - Underlying studies’ quality
  - Plural of “garbage” isn’t “data”

---

**STEPS**

1. Formulate a question
2. Systematic search of literature and unpublished studies
3. Select studies based on stated criteria
4. Select variable of interest
5. Analyze: pool the data and obtain result
SUPERIORITY VS. EQUIVALENCE

Not the same

Most of the Time

- Nearly all studies are superiority trials
  - Trying to reject null hypothesis of no difference by finding one
  - Is there an elephant in this room?
  - Failing to find a difference ≠ proving equivalence

- But what if you want to “prove” that drug A has the same effect as drug B?

An Equivalence Trial

- An equivalence trial lets you do that.
- First, we define a minimum clinically-significant difference $D_c$.
- Then we do our study and get a confidence interval for the difference between groups.
  - Does the confidence interval include zero and exclude $-D_c$ and $+D_c$?
- Typically requires large samples, because you need a small confidence interval.

Computer Knowledge

Not like Her
WHAT THIS IS FROM

CONTENT OUTLINE

JOINT COUNCIL ON ANESTHESIOLOGY EXAMINATIONS

AMERICAN BOARD OF ANESTHESIOLOGY

Revised - September, 2011

WHAT IT SAYS

3. Computer: Data Handling, Processing, and Analysis

a) Basic Computer Knowledge: Programs vs. Operating System, Computer Virus, Disk Or Central Processing Unit (CPU) Failure, Amplifiers, Microprocessors

ALL I’LL SAY

- Most statistical software wants data this way:
  - One person per line
  - Each variable gets a column
  - Numbers where appropriate, category designations otherwise

<table>
<thead>
<tr>
<th>Patient</th>
<th>Sex</th>
<th>Age (y)</th>
<th>Dose (mg)</th>
<th>Duration (h)</th>
<th>Max pain (0-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>42</td>
<td>10</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>44</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>35</td>
<td>14</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

THIS IS THE END

I promised there would be a bear