Mathematical Tools for Neural and Cognitive Science

Fall semester, 2020

Section 5:
Statistical Inference and Model Fitting

The sample average

\[ \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

What happens as \( N \) grows?

- Variance of \( \bar{x} \) is \( \sigma_x^2 / N \) (the “standard error of the mean”, or SEM), and so converges to zero [on board]
- “Unbiased”: \( \bar{x} \) converges to the true mean, \( \mu_x = \mathbb{E}(x) \) (formally, the “law of large numbers”) [on board]
- The distribution \( p(\bar{x}) \) converges to a Gaussian (mean \( \mu_x \) and variance \( \sigma_x^2 / N \)): formally, the “Central Limit Theorem”

700 samples

sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23 -0.23 0.29]

true mean: [0 0.8]
true cov: [1.0 -0.25 -0.25 0.3]
Central limit for a uniform distribution...

10k samples, uniform density (sigma=1)

\( (u_1 + u_2) / \sqrt{2} \)

\( (u_1 + u_2 + u_3 + u_4) / \sqrt{4} \)

central limit for a binary distribution...

one coin

any of 16 coins

any of 64 coins

Classical “frequentist” statistical tests

Statistical Rethinking, Richard McElreath
Classical/frequentist approach - $z$

- In the general population, IQ is known to be distributed normally with
  - $\mu = 100$, $\sigma = 15$
- We give a drug to 30 people and test their IQ
- $H_1$: NZT improves IQ
- $H_0$ ("null"): it does nothing

Test statistic

- We calculate how far the observed value of the sample average is away from its expected value.
- In units of standard error.
- In this case, the test statistic is

$$z = \frac{\bar{x} - \mu}{SE} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

- Compare to a distribution, in this case $z$ or $N(0,1)$

Does NZT improve IQ scores or not?

<table>
<thead>
<tr>
<th>Reality</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Correct</td>
<td>Miss</td>
</tr>
<tr>
<td>No</td>
<td>Type I error</td>
<td>Correct</td>
</tr>
<tr>
<td>Significant?</td>
<td>Yes</td>
<td>$\alpha$-error</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$\beta$-error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>False alarm</td>
</tr>
<tr>
<td>Correct</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The z-test

- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $\bar{x} = 108.3$ (Sample mean)
- $z = (\bar{x} - \mu)/SE = (108.3-100)/SE$ (Standardized score)
- $SE = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- Error bar/CI: $\pm 2$ SE
- $z = 8.3/2.74 = 3.03$
- $p = 0.0012$
- Significant?
- One- vs. two-tailed test

What if the measured effect of NZT had been half that?

- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $\bar{x} = 104.2$ (Sample mean)
- $z = (\bar{x} - \mu)/SE = (104.2-100)/SE$
- $SE = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- $z = 4.2/2.74 = 1.53$
- $p = 0.061$
- Significant?

Significance levels

- Are denoted by the Greek letter $\alpha$.
- In principle, we can pick anything that we consider unlikely.
- In practice, the consensus is that a level of 0.05 or 1 in 20 is considered as unlikely enough to reject $H_0$ and accept the alternative.
- A level of 0.01 or 1 in 100 is considered “highly significant” or “really unlikely”.
Common misconceptions

Is “Statistically significant” a synonym for:
- Substantial
- Important
- Big
- Real

Does statistical significance gives the
- probability that the null hypothesis is true
- probability that the null hypothesis is false
- probability that the alternative hypothesis is true
- probability that the alternative hypothesis is false

Meaning of $p$-value. Meaning of CI.

Student’s $t$-test

- $\sigma$ not assumed known
- Use $s^2 = \frac{\sum (x_i - \bar{x})^2}{N-1}$
- Why $N-1$? $s$ is unbiased (unlike ML version), i.e., $\mathbb{E}(s^2) = \sigma^2$
- Test statistic is $t = \frac{\bar{x} - \mu_0}{s / \sqrt{N}}$
- Compare to $t$ distribution for CIs and NHST
- “Degrees of freedom” reduced by 1 to $N-1$

The $t$ distribution approaches the normal distribution for large $N$
The z-test for binomial data

• Is the coin fair?
• Lean on central limit theorem
• Sample is \( n \) heads out of \( m \) tosses
• Sample mean: \( \hat{p} = \frac{n}{m} \)
• \( H_0: p = 0.5 \)
• Binomial variability (one toss): \( \sigma = \sqrt{pq} \), where \( q = 1 - p \)
• Test statistic: \( z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/m}} \)
• Compare to \( z \) (standard normal)
• For CI, use \( \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q}/m} \)

Many varieties of frequentist univariate tests

• \( \chi^2 \) goodness of fit
• \( \chi^2 \) test of independence
• Test a variance using \( \chi^2 \)
• \( F \) to compare variances (as a ratio)
• Nonparametric tests (e.g., sign, rank-order, etc.)

Lack of correlation is favored in \( N>3 \) dimensions

Null Hypothesis: Distribution of normalized dot product of pairs of Gaussian random vectors in \( N \) dimensions:

\[
1 - d^2 \left( \frac{N-3}{2} \right)
\]
Estimation of model parameters (outline)

- How do I compute estimated values from data?
- How “good” are my estimates?
- How well does my model explain data to which it was fit? Other data (prediction/generalization)?
- How do I compare models?

Estimation

- An “estimator” is a function of the data, intended to provide an approximation of the “true” value of a parameter.
- Traditionally, one evaluates estimator quality in terms of error mean (“bias”) and error variance (note: $\text{MSE} = \text{bias}^2 + \text{variance}$).
- Traditional statistics aims for an unbiased estimator, with minimal variance (“MVUE”).
- More nuanced contemporary view: trade off the bias and variance, through model selection, “regularization”, or Bayesian priors.

The maximum likelihood estimator (MLE)

Sample average is appropriate when one has direct measurements of the thing being estimated. But one may want to estimate something (e.g., a model parameter) that is indirectly related to the measurements…

Natural choice: assuming a probability model $p(\mathbf{x} | \theta)$ find the value of $\theta$ that maximizes this “likelihood” function:

$$\hat{\theta}(\mathbf{x}_n) = \arg \max_{\theta} \prod_{n} p(\mathbf{x}_n | \theta) = \arg \max_{\theta} \sum_{n} \log p(\mathbf{x}_n | \theta).$$
Example: Estimate the bias (probability of heads) of a coin, by observing some samples.

\[ p(H = 1, T = 0 | \theta) \]
\[ p(H = 0, T = 1 | \theta) \]
Likelihoods, \( p(H, T | \theta) \)

Convergence

Example ML Estimators - discrete

Binomial:  
\[
p(H | N, \theta) = \binom{N}{H} \theta^H (1 - \theta)^{N-H} \]
\[
\hat{\theta} = \frac{H}{N}
\]

Poisson:  
\[
p(k; \theta) = \frac{\theta^k e^{-\theta}}{k!}
\]
\[
\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} k_n
\]
Example ML Estimators - Continuous

Gaussian: \[ p \left( \{ x_n \} | \mu, \sigma \right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \]

\[ \hat{\mu} = \frac{\sum x_n}{N} \quad \hat{\sigma}^2 = \frac{\sum (x_n - \hat{\mu})^2}{N} \] (Note: this is biased!)

Uniform: \[ p \left( \{ x_n \} | \theta \right) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases} \]

\[ \hat{\theta} = \max_{n} \{ x_n \} \]

Properties of the MLE

• In general, the MLE is asymptotically unbiased and Gaussian, but can only rely on these if:
  - the likelihood model is correct
  - the MLE can be computed
  - you have lots of data

• Estimates of confidence:
  • SEM (relevant for estimates of mean)
  • inverse second deriv of NLL (multi-D: “Hessian”)
  • simulation (of estimates by sampling from \( p(x|\hat{\theta}) \))
  • bootstrapping (resample from the data, with replacement)

Bootstrapping

• “The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps” [Adventures of Baron von Munchhausen, by Rudolph Erich Raspe]

• A (re)sampling method for computing estimator distribution (incl. stdev error bars or confidence intervals)

• Idea: instead of looking at distribution of estimates across repeated experiments, look across repeated resampling (with replacement) from the existing data (“bootstrapped” data sets)
Bayesian Inference

\[ p(\theta | \text{data}) = \frac{p(\text{data} | \theta) p(\theta)}{p(\text{data})} \]

"Posterior" \hspace{1cm} "Likelihood" \hspace{1cm} "Prior"
Example: Posterior for coin

Infer whether a coin is fair by flipping it repeatedly here. \( x \) is the probability of heads (50% is fair) \( y_{1..n} \) are the outcomes of flips.

Consider three different priors:
- suspect fair
- suspect biased
- no idea

\[
\begin{align*}
\text{x likelihood (heads)} & = \text{posterior} \\
\text{X likelihood (heads)} & = \text{new posterior}
\end{align*}
\]
previous posteriors

X likelihood (tails)

= new posterior

Posteriors after observing 75 heads, 25 tails

→prior differences are ultimately overwhelmed by data

Bayesian confidence intervals

PDFs

CDFs, and 95% confidence intervals
MAP estimation - Gaussian case

For measurements with Gaussian noise, and assuming a Gaussian prior, posterior is Gaussian.

- MAP estimate is a weighted average of prior mean and measurement
- posterior is Gaussian, allowing sequential updating
- explains “regression to the mean”, as shrinkage toward the prior

MAP estimation - Gaussian case

Completing the square shows that this posterior is also Gaussian, with

\[
\frac{1}{\sigma^2} = \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2}
\]

\[
\mu = \left( \frac{y}{\sigma_y^2} + \frac{\mu_x}{\sigma_x^2} \right) / \left( \frac{1}{\sigma_y^2} + \frac{1}{\sigma_x^2} \right)
\]

The average of \( y \) and \( \mu_x \), weighted by inverse variances (a.k.a. “precisions”!)

MAP with Gaussians

\[
y = x + n, \quad x \sim N(\mu_x, \sigma_x), \quad n \sim N(0, \sigma_n)
\]

\[
p(x|y) \propto p(y|x)p(x)
\]

\[
e^{-\frac{1}{2} \left[ \frac{1}{\sigma_x^2} (x - \mu_x)^2 \right]} e^{-\frac{1}{2} \left[ \frac{1}{\sigma_n^2} (x - y)^2 \right]}
\]

\[
e^{-\frac{1}{2} \left[ \frac{1}{\sigma_x^2} (x - \mu_x)^2 \right] + \left( \frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2} \right) (x - y)^2}
\]

Two noisy measurements of the same variable:

\[
y_1 = x + n_1 \quad x \sim N(0, \sigma_x)
\]

\[
y_2 = x + n_2 \quad n \sim N(0, \sigma_n), \text{ independent}
\]

Joint measurement distribution:

\[
\tilde{y} \sim N(\tilde{0}, \sigma_y^2 I + \sigma_x^2 I)
\]

Least-squares regression

\[
\hat{\beta} = \text{arg min}_{\beta} \| y_2 - \beta y_1 \|^2
\]

\[
= \frac{\sigma_y^2}{\sigma_y^2 + \sigma_x^2}
\]

\[
E(y_2|y_1) = \hat{\beta} y_1
\]

“regression to the mean”
Regression to the mean

“Depressed children treated with an energy drink improve significantly over a three-month period. I made up this newspaper headline, but the fact it reports is true: if you treated a group of depressed children for some time with an energy drink, they would show a clinically significant improvement….”

“It is also the case that depressed children who spend some time standing on their head or hug a cat for twenty minutes a day will also show improvement.”

- D. Kahneman

The hierarchy of statistical estimators

- Maximum likelihood (ML): \( \hat{x}(\vec{d}) = \arg \max_x p(\vec{d}|x) \)
- Maximum a posteriori (MAP): \( \hat{x}(\vec{d}) = \arg \max_x p(x|\vec{d}) \) (requires prior, \( p(x) \))
- Bayes estimator (general): \( \hat{x}(\vec{d}) = \arg \min_{\hat{x}} \mathbb{E} \left( L(x, \hat{x}) \bigg| \vec{d} \right) \) (requires loss, \( L(x, \hat{x}) \))
- Bayes least squares (BLS): \( \hat{x}(\vec{d}) = \arg \min_{\hat{x}} \mathbb{E} \left( (x - \hat{x})^2 \bigg| \vec{d} \right) \) (special case, squared loss)

\[ \hat{x}(\vec{d}) = \mathbb{E} \left( x \bigg| \vec{d} \right) \]