PSYCH-GA 2240 – Fall 2023
Psychophysics
Mondays, 2-4, Room 159

Prof. Michael Landy – landy@nyu.edu

Often-used textbook:

Kingdom, F. A. A. & Prins, N. (2010). *Psychophysics: A Practical Introduction*. New York: Academic Press.

Other general references:

Baird, J. C. & Noma, E. (1978). *Fundamentals of Scaling and Psychophysics*. New York: Wiley.
Falmagne, J.-C. (1985). *Elements of Psychophysical Theory*. New York: Oxford.
Lu, Z.-L. & Dosher, B. A. (2014). *Visual Psychophysics: From Laboratory to Theory*. Cambridge, Mass.: MIT Press.

Software:

The Palamedes toolbox: http://www.palamedestoolbox.org

Psignifit: https://www.nip.uni-tuebingen.de/research/software/psignifit.html

Schedule and readings:

9/18: Introduction: Psychophysical tasks and procedures

9/25: Psychometric functions: how to fit, what to estimate, goodness of fit

10/2: Yes/no tasks, signal detection theory and the psychometric function

TUESDAY, 10/10: Adaptive procedures: Staircases, Quest, Pest, Ape, Psi and all that

10/16: Rating-scale methods and getting to high d', Interval bias, detection and identification

10/23: Techniques for fitting models: one, two or many parameters

10/30: Parameter estimation and confidence intervals

11/6: Controversies: Wichmann/Hill, Klein, Prins

11/13: Bayesian parameter estimation, Jeffries priors, marginalization

11/20: Model comparison I: Why use Bayesian inference? A cautionary tale

11/27: Model comparison II: Sampling methods

12/4: Model comparison III: Bayesian model comparison

12/11: Practical advice and packages for model comparison

9/18: Introduction: Psychophysical tasks and procedures

Reading: Kingdom & Prins, Ch. 1-3

References:

Farell, B. & Pelli, D. G. (1998). Psychophysical methods, or how to measure a threshold, and why. In Carpenter, R. H. S. & Robson, J. G. (Eds.), *Vision Research: A Practical Guide to Laboratory Methods* (pp. 129–136). New York: Oxford University Press.

Lu, Z.-L. & Dosher, B. A. (2014). *Visual Psychophysics: From Laboratory to Theory*. Cambridge, Mass.: MIT Press. Chapter 7.

Outline of the semester
Software packages for fitting
Textbook, readings
Overlap with other courses: Perception, Gureckis' Modeling course
Grading, exercises
How to solve exercises
- Palamedes
- Psignifit
- Read their code

Do it yourself (in Matlab, python, R, etc.)

I. Psychophysics
- Definition/Goals
- Type A vs. type B experiments, Sensitivity vs. appearance
- Detection vs. discrimination
- Psychometric function $P = f(x)$
- Ogive curve
  - 50% point, Point of subjective equality (PSE), Threshold
  - Slope

II. Psychophysical Methodology
- Concerns
  - Bias
  - Criterion
  - Attentiveness
  - Strategy
  - Artifactual cues
  - History of stimulation
  - Who controls stimulation
- Threshold methods
  - Method of adjustment
  - Method of (ascending/descending) limits
  - Method of constant stimuli (Yes-No)
  - Forced choice (2I2AFC, 3AFC, oddity, MAFC, ABX, etc.)

- Method of single stimuli
- Sequential testing (staircase methodologies)

Scaling methods

- Magnitude estimation, production, cross-modal matching
- Stevens power law
- Bisection (adjustment or forced-choice)
- Paired difference scaling (adjustment or forced-choice)
- Maloney's ML difference scaling procedure

9/25: Psychometric functions: how to fit, what to estimate, goodness of fit

Reading: Kingdom & Prins, Chs. 4, and 8.2.4, 8.3.1-8.3.3 (in the 1st edition) or 9.2.4, 9.3.1-9.3.3 (2nd edition). I'll also touch on material in Ch. 7 (2nd edition only)

References:

Carlin, B. P. & Lewis, T. A. (2009). *Bayesian Methods for Data Analysis* (3rd Ed.). New York: CRC Press. Section 2.5.1.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd Ed.). New York: CRC Press. Chapter 6.

Klein, S. A. (1985). Double-judgment psychophysics: problems and solutions. *Journal of the Optical Society of America A*, *2*, 1560–1585.

Lewandowsky, S. & Farrell, S. (2011). *Computational Modeling in Cognition* (Ch. 4). Washington, DC: Sage.

Lu, Z.-L. & Dosher, B. A. (2014). *Visual Psychophysics: From Laboratory to Theory*. Cambridge, Mass.: MIT Press. Chapter 10.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. fitting, sampling and goodness-of-fit. *Perception and Psychophysics*, *63*, 1293–1313.

Psychometric functions
- Basic constraints
  - Range of dependent variable
  - Log/linear scale, dB
  - Chance performance level
  - Linear vs. circular independent variable
  - Lapses
  - Goal is to estimate
    - Threshold as nominal performance level
    - Threshold as slope
    - Independent of lapses
    - PSE
- Models
  - Random threshold
  - Noise
    - Additive
    - Multiplicative (log law)
    - Multiple channels (Quick)
    - Uncertainty (Pelli)
- Parametric models
  - Probit/Cumulative normal: $P(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/2\sigma^2} dt$

,

where slope "β" = $1/\sigma$, "α" = μ

Log-normal: $P(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right)$

Note that μ and σ are in $\log x$ units

Logit/Logistic: $\frac{1}{1 + e^{-\beta(x-\alpha)}}$

Weibull: $1 - e^{-(x/\alpha)^\beta}$ for positive x only

Quick: $1 - 2^{-\left(R(x)\right)^\beta}$

Probability summation:

$$
\begin{aligned}
P(\text{detect}) &= 1 - P(\text{not detect}) \\
&= 1 - \prod_i P(\text{not detect in channel } i) \\
&= 1 - \prod_i (1 - P\left(\text{detect in channel } i)\right) \\
&= 1 - \prod_i \left(1 - \left(1 - 2^{-\left(R_i(x)\right)^\beta}\right)\right) \\
&= 1 - \prod_i 2^{-\left(R_i(x)\right)^\beta} \\
&= 1 - 2^{-\sum_i \left(R_i(x)\right)^\beta}
\end{aligned}
$$

Thus, probability summation is like a response summed over multiple channels (i.e., a vector length with Minkowski metric, Euclidean if $\beta = 2$)

Correction for guessing and lapses

What to estimate

Threshold or PSE (μ or α)

Slope (σ or β)

Fit criterion

Squared error (leading to χ^2 or *F* tests, variance accounted for, etc.

Maximum likelihood

Parameter vector (e.g., $\vec{\theta}(\alpha, \beta)$)

Likelihood $l(\vec{\theta}) = P(\text{data} \,|\, \vec{\theta})$

For psychophysical data:

Condition i, test at level x_i, data are n_i correct out of m_i trials

Assume independent trials, stable performance

Choose $\vec{\theta}$ that maximizes

$$
\begin{aligned}
l(\vec{\theta}) &= P(\text{data} \,|\, \vec{\theta}) \\
&= \prod_i P(\text{data}_i \,|\, \vec{\theta}) \\
&= \prod_i \binom{m_i}{n_i} \left(P_{\vec{\theta}}(x_i)\right)^{n_i} \left(1 - P_{\vec{\theta}}(x_i)\right)^{(m_i - n_i)}
\end{aligned}
$$

To avoid computer underflows, equivalently maximize

$$
\begin{aligned}
\log l(\vec{\theta}) &= \log P(\text{data} \,|\, \vec{\theta}) \\
&= \sum_i \log P(\text{data}_i \,|\, \vec{\theta}) \\
&= \sum_i \left[\log \binom{m_i}{n_i} + n_i \log P_{\vec{\theta}}(x_i) + (m_i - n_i) \log \left(1 - P_{\vec{\theta}}(x_i)\right) \right]
\end{aligned}
$$

Drop the first term because it does not depend on the parameters

Constrained parameters

Matlab: fmincon

Reparameterize

Half-line: exp/log

Finite interval: logistic $y = 1/(1 + e^{-x})$ and its inverse

$x = -\log\left((1/y) - 1\right)$

Bayesian methods (a topic for later in the semester)

Note: Maximum likelihood is the same as MAP with a flat prior

Goodness of fit

Basic χ^2 and why it's inappropriate

Deviance and goodness of fit

Saturated model

Likelihood ratio

Deviance $L = 2 \log \dfrac{l(M_{\text{saturated}})}{l(M_{\text{psychometric}}; \hat{\theta}}$

Nested hypothesis test

Degrees of freedom = # of extra parameters

= # levels - # parameters

Note: parameters must be "meaningful", i.e., "independent"

Alternative: bootstrapped deviance distribution

Deviance residuals (square root of deviance per datapoint):

$$
d_i = sgn(y_i - p_i) \sqrt{2 \left[m_i y_i \log \frac{y_i}{p_i} + m_i (1 - y_i) \log \frac{1 - y_i}{1 - p_i} \right]}
$$

Look at correlation between d_i and p_i to check the quality of the fit (e.g., to possibly reject the Weibull as a model for your data)

Look at correlation between d_i and k_i (the index for when that datapoint was collected, assuming levels were blocked, not mixed) for evidence of learning

Failure to fit and what to do about it

Homework: Please email me the results of the following and, if you like, the Matlab that generated them, all folded together as a single PDF. If you are only auditing, please do NOT send me anything! ;^) Due: October 16, 2PM

(1) Write Matlab code to simulate an observer in a 2AFC method of constant stimuli task. The observer is assumed to conform to a particular parametric form of the psychometric function (e.g., log-normal, Weibull, whatever), and you supply a fixed set of parameters (guessing=gamma, position alpha and slope beta, for now let lapses = 0). Generate a large set of sample psychometric functions (each of which consists of something like 40 trials at something like 5 or 7 levels).

(2) Use psignifit, Palamedes, or better yet, write your own Matlab code to fit that same parametric psychometric function to data, and run that fit on each simulated dataset.

(3) Plot the histogram of estimated parameters (or a 2-D contour plot of the 2-D histogram of α and β) and indicate the veridical value.

(4) For at least one dataset, fit ANOTHER parametric form (e.g., Weibull instead of logistic) and plot the two fit psychometric functions together to see where they are close and where they diverge.

10/2: Yes/no tasks, signal detection theory and the psychometric function

Reading: Kingdom & Prins, Ch. 6

References:

Green, D. M. & Swets, J. A. (1989). *Signal Detection Theory and Psychophysics*. Los Altos Hills, CA: Peninsula Publishing.

Lu, Z.-L. & Dosher, B. A. (2014). *Visual Psychophysics: From Laboratory to Theory*. Cambridge, Mass.: MIT Press. Chapter 8.

Macmillan, N. A. & Creelman, C. D. (2004). *Detection Theory: A User's Guide* (Chs. 1–2). New York: Psychology Press.

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. New York: Oxford.

Background: Thurstone

One-dimensional theory

- Signal and noise distributions
- Maximum likelihood approach, likelihood ratio
- Equal variance case
  - Hits, misses, false alarms, correct rejections
  - Criterion
  - Calculating the probabilities
  - Calculating sensitivity $d' = z(H) - z(FA)$ and criterion/bias β
  - Varied criterion: the isosensitivity or ROC curve
  - ROC/AOC/NOC (Barlow)/etc.
  - Noisy hard threshold and its ROC, high threshold theory, etc.
  - Optimal criterion
    - Optimality: maximum percent correct, maximum utility, etc.
    - Define
      - V_{Ys} to be the value of saying yes on a signal trial
      - V_{Ns} to be the cost of saying no on a signal trial
      - etc.

    $E(Y|x) = V_{Ys}P(s|x) - V_{Yn}P(n|x)$

    $E(N|x) = V_{Nn}P(n|x) - V_{Ns}P(s|x)$

    Say yes if

    $$E(Y|x) \geq E(N|x)$$

    That is, when

    $$\frac{P(s|x)}{P(n|x)} \geq \frac{V_{Nn} + V_{Yn}}{V_{Ns} + V_{Ys}}$$

    Use Bayes rule

    $$P(s|x) = \frac{P(x|s)P(s)}{P(x)}$$

    To derive

    $$\frac{P(s|x)}{P(n|x)} = \frac{P(x|s)}{P(x|n)}\frac{P(s)}{P(n)}$$

In words:

posterior odds = likelihood ratio × prior odds

Thus, say yes if

$$l(x) = \frac{P(x\,|\,s)}{P(x\,|\,n)} \geq \frac{P(n)}{P(s)} \frac{V_{Nn} + V_{Yn}}{V_{Ys} + V_{Ns}} = \beta$$

Equal utility, equal priors: $\beta = 1$

Effect of priors and payoffs

ROC slope as the ratio of the standard deviations

Gaussian assumption $N(\mu, \sigma)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Double probability paper and fitting

2IFC performance

Area under the ROC

For Gaussian case:

$$\begin{aligned} P(C) &= P(N(s, \sigma^2) > N(0,\sigma^2)) \\ &= P(N(s, \sigma^2) - N(0,\sigma^2) > 0) \\ &= P(N(s,2\sigma^2) > 0) \\ &= P(N(0,2\sigma^2) < s) \\ &= P\left(N(0,1) < \frac{s}{\sqrt{2}\sigma}\right) \\ &= P\left(N(0,1) < \frac{d'}{\sqrt{2}}\right) \end{aligned}$$

Hence, $P(C)$ is a cumulative normal

Area under the ROC and forced choice performance

ROCs from a single rating scale experiment

Unequal variance case

Maximum likelihood versus setting a criterion

ROC asymmetry

Multidimensional theory

Forced choice as two dimensions reduced to one ($\sqrt{2}$ factor)

Multivariate Gaussians and statistical decision theory

10/10: Adaptive procedures: Staircases, Quest, Pest, Ape, Psi and all that

Reading: Kingdom & Prins, Ch. 5

References:

Cornsweet, T. N. (1962). The staircase method in psychophysics. *American Journal of Psychology*, *75*, 485–491.

Findlay, J. M. (1978). Estimates on probability functions: A more virulent PEST. *Perception & Psychophysics*, *23*, 181–185.

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, *38*, 1861-1881.

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, *69*, 1763–1769.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics*, *49*, 227-229.

Kesten, H. (1958). Accelerated stochastic approximation. *Annals of Mathematical Statistics*, *29*, 41-59.

Kontsevich, L. L. & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*, 2729–2737.

Lesmes, L. A., Lu, Z.-L., Baek, J., Tran, N., Dosher, B. A. & Albright, T. D. (2015). Developing Bayesian adaptive methods for estimating sensitivity thresholds (d') in Yes-No and forced-choice tasks. *Frontiers in Psychology*, *6:1070*.

Lesmes, L. A., Lu, Z.-L., Tran, N. T., Dosher, B. A. & Albright, T. D. (2006). An adaptive method for estimating criterion sensitivity (d') levels in yes/no tasks. *Journal of Vision*, *6(6)*, 1097.

Lesmes, L. A., Jeon, S. t., Lu, Z.-L. & Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: the quick TvC method. *Vision Research*, *46*, 3160-3176.

Lesmes, L. A., Lu, Z.-L., Baek, J. & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: the quick CSF method. *Journal of Vision*, *10(3):17*, 1-21.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*, 467–477.

Lu, Z.-L. & Dosher, B. A. (2014). *Visual Psychophysics: From Laboratory to Theory*. Cambridge, Mass.: MIT Press. Chapter 11.

Macmillan, N. A. & Creelman, C. D. (2004). *Detection Theory: A User's Guide* (Ch. 8). New York: Psychology Press.

Owen, L., Browder, J., Letham, B., Stocek, G., Tymms, C. & Shvartsman, M. (2021). Adaptive nonparametric psychophysics. https://arxiv.org/abs/2104.09549 and https://aepsych.org/

Pentland, A. (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics*, *28*, 377–379.

Prins, N. (2013). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision*, *13(7):3*, 1-17.

Robbins, H. & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400-407.
Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America*, *49*, 505–508.
Taylor, M. M. & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, *41*, 782–787.
Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, *35*, 2503–2522.
Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, *17(3):10*, 1-27.
Watson, A. B. & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113–120.
Watt, R. J. & Andrews, D. P. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, *1*, 205–214.
Wetherill, G. B. (1966). Sequential estimation of points on quantal response curves. In *Sequential Methods in Statistics* (pp. 171–227). London: Methuen.
Wetherill, G. B. & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, *18*, 1–10.

Staircase procedures
Concerns
- Computation during trials
- Efficiency/sweat factor/number of trials/trial placement
- Subject fatigue (boredom if too easy, frustration if too hard)
- Stationarity
- Finger errors
- Desired estimates: $L_{.5}$, L_p, slope
- Sequential dependencies, interleaved staircases
- Estimation bias
- Correction for guessing and for finger errors

Assumptions
- Monotonic
- Threshold approximately known
  - Slope β known or approximately known
  - Parametric form of f
  - Stationary
  - Independent trials (interleaving)
- Basics
  - How to place trials
  - When to stop
  - How to estimate parameters

Procedures
- Robbins/Monro, Kesten

$$x_{n+1} = x_n + \frac{c}{c}\left(p - y_n(x_n)\right)$$

to estimate L_p, biased away from 50%

Up-Down (Dixon/Mood, Cornsweet)

Transformed Up-Down (Levitt/Weatherill)

1-up-2-down, 2-up-1-down, 1-up-3-down, etc.

Halve stepsize every other turnaround and restart at current threshold estimate

Notion of a transformed response curve

Weighted Up-Down (Kaernbach), Transformed/weighted (García-Pérez)

PEST (Taylor & Creelman, Findlay, Pentland)

Wald test to change levels, changes in step size to deal with closeness and distance from correct spot, stop at minimum step size

APE (Watt & Andrews)

Method of constant stimuli for blocks of trials, then fit previous 2 blocks and choose a new set of levels ranging over ± 1.35 SD with momentum based on prior change

QUEST (Watson & Pelli), β as a constant for log scaled stimulus strength

Sweat factor (Taylor/Creelman): $K = N\sigma_{est}^2$

Ideal sweat factor = $p(x)q(x)/\left(\dfrac{dP_T}{dx}\right)^2$

Don't know *T* so use maximum a posteriori

Maximize $P(T\,|\,D) = \dfrac{P(D\,|\,T)P(T)}{P(D)}$, by Bayes rule

So, maximize Quest function $Q(T) = \log P(T) + \log P(D\,|\,T)$

Assume independent trials, so

$$\log P(D\,|\,T) = \log\left(\prod_i P(R_i\,|\,x_i, T)\right) = \sum_i \log P(R_i\,|\,x_i, T)$$

Addend is either $\log P_T(x) = \log \Psi(x - T)$

or $\log\left(1 - P_T(x)\right) = \log\left(1 - \Psi(x - T)\right)$

so precompute these and accumulate over trials

Log likelihood $L(T) = Q(T) - Q_0(T)$

where $Q_0(T) = \log P(T)$, the log of the *prior*

Stop based on a likelihood ratio test (a χ^2 test)

PSI method (Kontsevich & Tyler)

Estimates both α and β

Assumes independent priors on each

Does a Bayesian update after each trial

Chooses a level to test such that the expected entropy of the posterior after that trial is minimized, where entropy is:

$$H = -\int\int p(\alpha,\beta)\log p(\alpha,\beta)\,d\alpha\,d\beta$$

        - Estimate is the mean of the posterior
    - Psi-marginal method
    - Quick methods: q-YN, q-TvC, q-CSF
    - QUEST+ as a generalization of all of these (Mathematica, Matlab & Python)
    - AEPsych
- When to stop
    - *N* trials
    - *N* turnarounds
    - Given standard error of the estimate
- Estimation
    - Probit analysis
    - Midrun estimates
    - Maximum likelihood
    - Final values
    - Minimum χ^2
- Tips
    - Plot staircases: trial vs. level
    - Plot psychometric function with symbol area proportional to number of trials

10/16: Rating-scale methods and getting to high d’,
interval bias, history effects, detection and identification

Reading:

Yeshurun, Y., Carrasco, M. & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*, 1837–1851 [and corrigendum].

References:

Busse, L., Ayaz, A., Dhruv, N. T., Katzner, S., Saleem, A. B., Schölvinck, M. L., Zaharia, A. D. & Carandini, M. (2011). The detection of visual contrast in the behaving mouse. *Journal of Neuroscience*, *31*, 11351-11361.

Fründ, I., Wichmann, F. A. & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14(7):9*, 1-16.

Klein, S. A. (1985). Double-judgment psychophysics: problems and solutions. *Journal of the Optical Society of America A*, *2*, 1560–1585.

Macmillan, N. A. & Creelman, C. D. (2004). *Detection Theory: A User’s Guide* (Chs. 3, 5–7, 9). New York: Psychology Press.

Wickens, T. D. (2001). *Elementary Signal Detection Theory* (Chs. 5, 6.3 & 7). New York: Oxford.

Thurstone and d' scaling

How to summarize discriminability (or detectability) when noise depends on signal
- d_a, area under the ROC, and its variants

The relationship of detection (of A or of B) and discrimination (of A vs. B): univariate vs. independent vs. similar stimuli

Identification of multiple stimulus levels using the M-AFC task (Klein)

Klein (1985): 2x2 task
- Single knob task and monopolar and bipolar mechanisms
- I/D ratio and available mechanisms
- 2 knob summation task (blank, S1, S2, S1+S2)
- Usefulness of using a rating-scale task
- Criterion bias vs. correlated noise vs. inhibition vs. fluctuating attention.

2IFC vs. Yes-no
- In forced choice, the subject gets two noisy samples (x_1, x_2) which can be drawn from (S,N) or (N,S), whereas in yes-no, the subject gets one sample x drawn either from S or N. In standard yes-no, we get a hit rate and false-alarm rate and estimate what I now notate as d'_{YN}. In 2IFC, we can treat (S,N) as the “signal” and (N,S) as “target” and compute a hit rate (proportion of correct on interval-1 trials, PC_1) and a false-alarm rate (proportion of incorrect on interval-2 trials, 1- PC_2), resulting in d'_{FC}.
- Case 1: constant noise, no interval bias:

$$d'_{FC} = \sqrt{2}d'_{YN} = z(PC_1) + z(PC_2) = 2z(PC_{2IFC}).$$

Case 2: interval bias, it's still true that $d'_{FC} = \sqrt{2} d'_{YN} = z(PC_1) + z(PC_2) = \dfrac{d'_1 + d'_2}{\sqrt{2}}$,

an interval-bias-corrected estimate of d'. It is incorrect to ignore interval bias, i.e., to set $d'_{YN} = \sqrt{2} z(PC_{2IFC})$.

Case 3: possible interval bias *and* the noise for S+N differs from the noise for N. For this case, as with yes-no, a single criterion (or in the 2-d (x_1, x_2) space, a single criterion line) is suboptimal. The optimal observer, in fact, uses *two* criterion lines, or four decision regions. One criterion line is what Yeshurun et al. call the difference observer (a criterion on $x_1 - x_2$). The other flips the decision if $x_1 + x_2$ is sufficiently small. However, if you ignore that subtlety (as does the Wickens book), since so few samples end up in that region, and assume noise SD is 1 and signal mean and SD are (μ_S, σ_S), then it's easy to show that $d'_{FC} = \dfrac{2\mu_S}{\sqrt{1 + \sigma_S^2}}$.

Possibly false assumptions underlying interpretation of 2IFC (Yeshurun et al., note that there are published errata):

Four false assumptions:

1) $p_1 = p_2$ (i.e., 2IFC is often biased)

2) $d'_1 = d'_2$ (i.e., the procedure affects sensitivity)

3) $d'_{FC} = \tau d'_1$ (where $\tau > 1$ depends on the two sensitivities d'_1 and d'_2, so 2IFC performance cannot be predicted from individual Yes-No performances)

4) $d'_{FC} = d'_{YN}$ (2IFC is not always more sensitive than Yes-No)

Estimating d' in the presence of history effects using a GLM

Busse et al.: decision variable with bias terms (to stay or switch) based on previous trial's success or failure

Fründ et al.: decision variable with bias terms based on previous *n* responses and actual stimulus values

Geometry and ideal-observer analysis of more complex tasks: Same-different, ABX, Oddity vs. 3AFC

Homework: Simulate datasets for 2AFC tasks with method of constant stimuli for observers without and with interval bias, with either constant noise or possibly with signal-dependent noise (as in some models of Weber's Law) and, if you are motivated to do so, history effects. Then, analyze the psychometric functions using the tools you developed last time. Things you can try: (1) plot psychometric functions using the d' formula that ignores interval bias, and the d' formula that corrects for bias. Note, here I am referring to a psychometric function with d' on the *y*-axis rather than percent correct. (2) Scatterplot the d' values against one another. (3) Fit the d' psychometric functions (think about what function makes sense to fit to these) calculated both ways.
(4) Scatterplot the estimates of the fit curves against one another. How large an interval bias is required for significant effects on d' estimation? Due 11/13, 2PM.

10/23: Techniques for fitting models: one, two or many parameters

References:

Lewandowsky, S. & Farrell, S. (2011). *Computational Modeling in Cognition* (Section 3.1). Washington, DC: Sage.

Numerical analysis
- Efficiency
- Accuracy
- Dealing with quantization (roundoff) errors and underflows

Example: Finding a zero (Newton's method)

Finding a minimum or maximum (e.g., maximum-likelihood estimation)
- Gradient descent
  - Convexity, multiple local minima
  - Random starting points
  - 1d, gradient in *n* dimensions $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n} \right)$
  - How to compute: discrete derivatives, e.g., $(-2x_n + x_{n-1} + x_{n+1})/\delta$
  - Matlab: DERIVEST/HESSIAN suite
- Gridding, variants with random jitter, etc.
- Nelder-Mead simplex method
- Simulated annealing
- Mostly for discrete models: Genetic algorithms
- Stochastic gradient descent
- Issues with stochastic error functions

Fancier methods: EM, MCMC (later!)

10/30: Parameter estimation and confidence intervals

References:

- Kärnbach, C. (2001). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics*, *63*, 1389-1398.
- Lewandowsky, S. & Farrell, S. (2011). *Computational Modeling in Cognition* (Section 5.1). Washington, DC: Sage.
- Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, *12(6):25*, 1-16.
- Sivia, D. S. & Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial* (2nd Ed.). Oxford, UK: Oxford University Press. Section 2.2.

How does one get an error bar on a parameter after a maximum-likelihood fit?

- SE across sessions
- SE across subjects (with a different meaning of error)
- Bootstrapping
  - Nonparametric and problems with adaptive methods
  - Parametric using visited levels
  - Parametric using same adaptive method
  - Problems with across-trial correlations in adaptive methods (Prins, Kärnbach)
    - Parametric and even non-parametric ML estimation depends on independent trials and responses
    - Adaptive methods, e.g., staircases, place stimuli on trial n+1 based on response on trial n, so are dependent in placement and in response
    - The result is slope bias: estimates are biased to be too steep. The bias is respectably high for 100 trials, and very high for 50 trials or fewer
    - The bias is not due to uneven trial placement as demonstrated by double-trial simulations (one for placement, a second for estimation)
    - The bias is due to re-test or test-at-all probability dependent on previous trial (examples of “do 2nd trial only if first is negative” and two-trial 1-up-1-down staircase visits level L-1 only if response at L was positive)
    - This is repaired if using an adaptive procedure that also places trials based on learning about slope (Kärnbach), and the problem returns when using adaptive procedures in the presence of lapses without explicitly trying to estimate them using the procedure (Prins)
- Maximum-likelihood vs. Bayesian approaches (from posterior, 2 classes hence)
- Curvature of the log-likelihood function: intuition
- Curvature as 2nd derivative
- Hessian matrix $H = \left[\dfrac{\partial^2 \log L(\theta \mid y)}{\partial\theta_i \partial\theta_j}\right]$
- Hessian and Fisher information

Covariance matrix $= H^{-1}$
Square root of diagonal elements of H^{-1} are standard errors
Correlated parameters, effective number of parameters (see DIC, later)
Off-diagonal elements give covariance of parameters
Like a Taylor series approximation at the mode. This is a quadratic approximation to the log-likelihood, thus a Gaussian approximation to the likelihood itself. For a posterior, this effectively approximates the posterior with a normal.

Note: maximum-likelihood estimates need not be unbiased

Example: 1-d Gaussian. The data are $x_1, x_2, \cdots, x_N$

$$\log L(\mu, \sigma \mid \vec{x}) = \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2}\right)$$

$$= \sum_{i=1}^{n}\left(-\log\left(\sqrt{2\pi}\sigma\right) - (x_i-\mu)^2/2\sigma^2\right)$$

$$= -N \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

At the maximum, the partial derivatives vanish:

$$0 = \frac{\partial \log L}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_{i=1}^{N} -2(x_i-\mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - N\mu\right)$$

$$0 = \frac{\partial \log L}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_{i=1}^{N} -2\left(x_i-\mu\right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - N\mu\right)$$

From which we derive $\hat{\mu} = \sum_{i=1}^{N} x_i / N = \bar{x}$, the usual sample mean

$$0 = \frac{\partial \log L}{\partial \sigma} = \frac{-N}{\sqrt{2\pi}\sigma}\sqrt{2\pi} - \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{2}(-2)\sigma^{-3}$$

$$= \frac{-N}{\sigma} + \sigma^{-3}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

From which we derive

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}$$

Note that this is the biased version of a sample-variance estimate

FYI: Note that the maximum-likelihood estimate of σ gives effectively the same answer as the maximum-likelihood estimate of σ^2, because there is no re-parameterizing of a distribution of σ for ML estimation.

11/6: Controversies: Wichmann/Hill, Prins

References/Readings:

Kärnbach, C. (2001). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics*, *63*, 1389-1398.

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*. *12(5):25*, 1–16.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. fitting, sampling and goodness-of-fit. *Perception and Psychophysics*, *63*, 1293–1313.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: II. bootstrap-based confidence intervals and sampling. *Perception and Psychophysics*, 63, 1314–1329.

Wichmann & Hill's two-paper sequence introduces the theory behind *psignifit*:

- Maximum-likelihood fits, MOCS, several stimulus-level regimes tested
- Include lapse rate, constrained to lie between 0 and 6%
- Without lapse rate, lapses are confounded with lower slope/higher threshold
  - leading to bias: Errors near $p = 0$ or 1 get huge weight, so a single error at high stimulus strength forces fit away from 1.0 at that stimulus level (and the same for a psychometric function asymptote at $p = 0$)
- Goodness of fit:
  - Pearson's χ^2 goodness-of-fit test vs.
  - χ^2 deviance test (nested hypothesis test vs. saturated model) vs.
  - *p*-value from bootstrapping deviance
    - Pearson's isn't optimized at ML parameters and is useless for model comparison
    - Deviance should be χ^2 with d.o.f. equal to the number of MOCS levels minus the number of curve parameters, but often isn't, so use Monte Carlo to get the deviance distribution. That is, asymptotic deviance χ^2 *p*-values can be quite wrong
    - reminder: deviance $= 2(\log L(\text{saturated}) - \log L(\text{fit}))$
  - overdispersion due to wrong model (e.g., wrong *F*)
- Precision estimated by bootstrapped WCI_{68}, they recommend parametric
  - bootstrap, but WCI_{68} based on $\hat{\theta}$ can be biased (too small) compared to one based on θ, i.e., the bootstrap bridging assumption (that the size of the CI is stable near θ) is often incorrect. They suggest using a 9-point grid around $\hat{\theta}$ of width based on WCI_{68} to check for potential bias and possibly, conservatively, substitute the max ($MWCI_{68}$). Choosing a different form of *F* than that which generated the data can result in huge differences in precision.

Prins's failed replication

- 3D log-likelihood plots
- A high lapse rate will be affected little by a single actual lapse whereas a low/zero lapse rate and a single lapse will result in a much shallower estimate of slope

- If no level is included for which the predicted p(yes) is near asymptote, get a ridge in the log-likelihood plot with slope trading off with lapse rate and the estimated lapse rate bounces back and forth between the ends of its constrained range independent of the generating lapse rate
- If include such a high level and if get 100% yes at that level then the estimate of the lapse rate will be zero. But, if the lapse rate is high, then you will get errors at that level and again fits will bounce between high lapse rate/high percentage correct and low lapse rate/low slope

Kärnbach points out that with staircases there is a bias in the slope estimate because the choice of visited levels depends on the data. The psi method, which simultaneously estimates threshold and slope, improves on this

Prins: if you don't design the method to estimate the lapse rate, there will be bias. Therefore, he suggests adding a very high stimulus level to pin the lapse rate (either using its percentage correct to estimate the lapse rate separately (assuming the underlying psychometric function equals 1 there) before fitting the rest of the curve, or doing both jointly)

11/13: Bayesian parameter estimation, Jeffries priors, marginalization

Reading: Kingdom & Prins, 4.3.3.2

References:

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer. Chapters 9 and 11.

Carlin, B. P. & Lewis, T. A. (2009). *Bayesian Methods for Data Analysis* (3rd Ed.). New York: CRC Press. Sections 2.2 and 5.2.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd Ed.). New York: CRC Press. Sections 1.3, 2.4, 2.8, 3.1-3.5 and Chapter 13.

Leonard, T. & Hsu, J. S. J. (1999). *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. New York: Cambridge.

Schütt, H. H., Harmeling, S., Macke, J. H. & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric function for (potentially) overdispersed data. *Vision Research*, *122*, 105-123.

Sivia, D. S. & Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial* (2nd Ed.). Oxford, UK: Oxford University Press. Chapters 2-3.

Bayes, take II

Given vector of data *y* and of unknown parameters θ associated with model *M*

Posterior $p(\theta \mid y, M) = \dfrac{p(y \mid \theta, M)p(\theta \mid M)}{p(y \mid M)}$, where

Prior predictive distribution is $p(y \mid M) = \int p(y \mid \theta, M)p(\theta \mid M)d\theta$ normalizes the posterior but can be ignored for determining best value of θ

Posterior predictive distribution, for sampling/bootstrapping/etc., is

$$\begin{aligned} p(\tilde{y} \mid y, M) &= \int p(\tilde{y}, \theta \mid y, M)d\theta \\ &= \int p(\tilde{y} \mid \theta, y, M)p(\theta \mid y, M)d\theta \\ &= \int p(\tilde{y} \mid \theta, M)p(\theta \mid y, M)d\theta \end{aligned}$$

Can report MAP, posterior mean, percentiles, shortest error bar (asymmetric error bars either way)

Posterior is a compromise between the prior and the data. Posterior's variance is always smaller than prior's variance

Types of priors

Flat prior. Depends on parameterization: If $\phi = h(\theta)$ then $p(\phi) = p(\theta) \mid h'(\theta) \mid^{-1}$ Example: If prior on σ is flat, corresponding distribution on σ^2 isn't.

Proper vs. improper (e.g., flat on infinite domain) priors. Improper priors can lead to proper posteriors, but be careful.

Conjugate priors (especially convenient)

Definition: a prior is *conjugate* for a given likelihood if it results in a posterior from the same distributional family

Example I: Beta distribution for coin flips

Assuming a flat prior for the m $p(\theta \,|\, m, n) \propto p(n \,|\, \theta, m) \propto \theta^n (1-\theta)^{m-n}$oment: , i.e., $\mathrm{Beta}(n+1, m-n+1)$

Hence, conjugate prior is $\theta \sim \mathrm{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ which acts like $\alpha - 1$ and $\beta - 1$ extra coin flips. It's flat when $\alpha = \beta = 1$.

Example II: Normal distribution

Conjugate prior is normal, $p(y) \propto \exp\left(-(y-\theta)^2/2\sigma^2\right)$ $p(\theta) \propto \exp\left(-(\theta-\mu_0)^2/2\tau_0^2\right)$ and hence

$$p(\theta \,|\, y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

$$\propto \exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right),$$

where $\dfrac{1}{\tau_1^2} = \dfrac{1}{\tau_0^2} + \dfrac{1}{\sigma^2}$ and $\mu_1 = \dfrac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$, i.e., the usual optimal cue integration equations. For multiple observations, you get the same answer except substituting $\bar{y}$ and σ^2/n.

Example III: Exponential families of distributions:

$p(y \,|\, \theta) = h(y) g(\theta) \exp\left(\eta(\theta) T(y)\right)$

which includes many standard distributions: Normal, Bernoulli, binomial, Poisson, exponential, Weibull, Laplace, chi-squared, log-normal, gamma, beta, etc.

After N observations y_i,

$$p(\vec{y} \,|\, \theta) = \prod_i h(y_i) g(\theta)^N \exp\left(\eta(\theta) \sum_i T(y_i)\right),$$ i.e.,

$\sum_i T(y_i)$ is a sufficient statistic (like the sample average).

The conjugate prior is $p(\theta \,|\, \chi, \nu) \propto f(\chi, \nu) g(\theta)^\nu \exp(\nu \theta^T \chi)$, , which, when combined with the N samples, yields a posterior of the form

$$p(\theta \mid y, \chi, \nu) \propto g(\theta)^{\nu+N} \exp\left(\theta^T\left(\sum_i T(y_i) + \nu\chi\right)\right).$$

That is, the prior acts like a set of ν pseudo-observations, each of which has sufficient statistic χ.

Informative priors (e.g., using knowledge of the population) vs. noninformative

Jeffreys priors – use a rule so that after a change of parameterization you still get a Jeffreys prior. The resulting constraint is to have the prior proportional to the square root of the Fisher information (of the data concerning the parameter), thus: $p(\theta) \propto \sqrt{J(\theta)} = \sqrt{E_y\left(\frac{d^2 \log p(y \mid \theta)}{d\theta^2} \mid \theta\right)}$

Example I: The Jeffreys prior for a mean or any location parameter is flat

Example II: The Jeffreys prior for σ or any scale parameter is $1/\sigma$

Both of these are improper priors if over an infinite range

Example III: Binomial. Jeffreys prior is $\mathrm{Beta}(1/2, 1/2)$ (i.e., not flat)

Maximum entropy priors given constraints (e.g., normal is MaxEnt given μ and σ)

Marginalizing and why

Example: Schütt et al. (2016) – psignifit 4

Nuisance parameters

Example: Suppose your model is a normal distribution, but you only care about μ not σ. You carry out your experiment and determine the joint posterior distribution $p(\mu, \sigma \mid y)$. You could report the value of μ corresponding to the joint MAP estimate, i.e., the pair $(\hat{\mu}, \hat{\sigma})$ that has maximal posterior probability. But, that effectively gives too much credence to the particular value of $\hat{\sigma}$ in which you have little belief. So, it makes more sense to integrate out this "nuisance parameter": $p(\mu \mid y) = \int p(\mu, \sigma \mid y) d\sigma = \int p(\mu \mid \sigma, y) p(\sigma \mid y) d\sigma$, which can be computed analytically, numerically, or using sampling procedures such as MCMC (see: next week).

Finding the posterior or marginal mode

Conditional maximization: split parameter set into mutually exclusive subsets. One subset at a time, maximize posterior for that subset while holding the others constant. Iterate. If it's one single parameter at a time, you can find the local maximum from the current value using Newton-Raphson (approximating the curve as a quadratic and jumping to its maximum), using numerical estimates of the first and second derivative

EM (expectation/maximization) algorithm

Distinguish *parameters* from *latent variables*, where the latter might be missing data (for which guesses can be made based on the parameters) or hidden, unobservable variables

Most useful when the log likelihood cannot be factored when both parameters and latent variables are unknown (e.g., the equation contains a log of a sum), but is simple to factor and maximize when either the latent variables or parameters are fixed. So, it's like conditional maximization in the sense of holding one *set* fixed at a time and iterating.

Example: Gaussian mixture model (see Bishop Fig. 9.5, and Ng teaching notes)

Multivariate data $X = \overrightarrow{x_1}, \overrightarrow{x_2}, \cdots, \overrightarrow{x_N}$

Model: mixture of K multivariate Gaussians $N(\mu_k, \Sigma_k)$ with probability π_k

Latent variables are z_{nk}, which is an indicator variable, set to one if $\vec{x}_n$ belongs to cluster (Gaussian) k

Simple non-parametric algorithm: *K*-means clustering (Bishop Fig. 9.1)

Start: Pick K (possibly arbitrary) means μ_k

Iterate:

1. Assign each $\vec{x}_n$ to the nearest μ_k
2. Recompute each μ_k as the mean of the $\vec{x}_n$ assigned to it

EM (Expectation-maximization) algorithm applied to Gaussian mixtures is like *K*-means except: estimates both the means and covariances of each cluster as it proceeds, and does a soft assignment of each data point to the clusters rather than picking a single cluster

Issues for maximum-likelihood

Singularities (shrinking around a data point), so infinite likelihood

Identifiability (permuting the clusters), so multiple identical peaks

EM Gaussian-mixture algorithm (Bishop Fig. 9.8)

Start: Pick initial values of $\left\{\mu_k, \Sigma_k, \pi_k\right\}$

Iterate:

1. E Step: Evaluate the "responsibilities" using current parameters: $\gamma(z_{nk}) = \dfrac{\pi_k p(\overrightarrow{x_n} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{N} \pi_j p(\overrightarrow{x_n} \mid \mu_j, \Sigma_j)}$
2. M Step: Re-estimate the parameters using current responsibilities:

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \overrightarrow{x_n}$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\overrightarrow{x_n} - \mu_k^{\text{new}})(\overrightarrow{x_n} - \mu_k^{\text{new}})^T$$

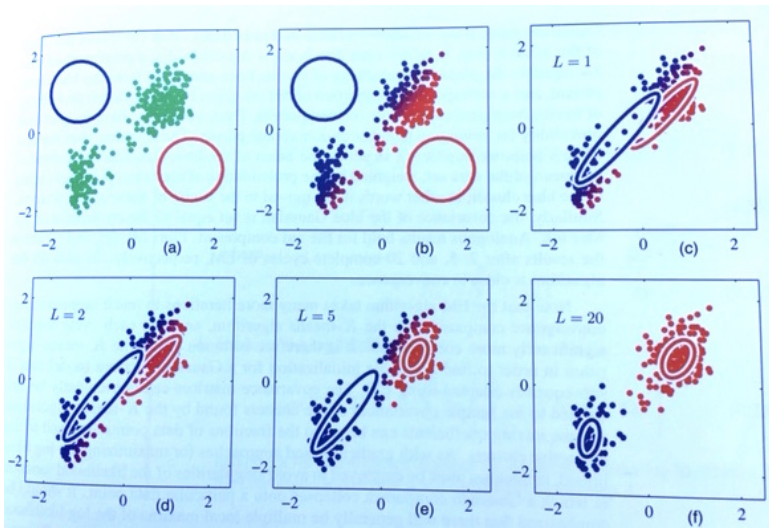
$\pi_k^{\text{new}} = \dfrac{N_k}{N}$, where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

3. Evaluate log likelihood

$$\log p(X \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k p(\overrightarrow{x_k} \mid \mu_k, \Sigma_k) \right\}$$

and check for convergence (parameters or log likelihood stopped changing)



Bishop Fig. 9.8: EM algorithm for Gaussian mixture

General EM: (1) Expectation: Update the estimates of the distribution of latent variable values ($\vec{z}$) conditional on the current estimate of the parameters θ^{old}, i.e., compute the expected sufficient statistics.
(2) Maximize the posterior density to determine a new estimate of the parameters θ.

More specifically:

E-step: Evaluate $p(\vec{z} \mid X, \overrightarrow{\theta^{\text{old}}})$

M-step: Pick new parameters to maximize based on the just-computed distribution of $\vec{z}$:

$$\overrightarrow{\theta^{\text{new}}} = \arg\max \sum_{\vec{z}} p(\vec{z} \mid X, \overrightarrow{\theta^{\text{old}}}) \log p(X, \vec{z} \mid \theta)$$

Standard error from percentiles of the marginal of the posterior

11/20-12/11: Model checking and comparison: Goodness of fit vs. overfitting, likelihood ratio, cross-validation, AIC, BIC, DIC, Bayes factor

Reading: Kingdom & Prins, Ch. 8 (1st edition) or 9 (2nd edition)

References:

Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer. Sections 1.3, 3.4, 4.4.1 and Chapters 8 and 11.

Bretthorst, G. L. (1996). An introduction to model selection using probability theory as logic. In Heidbreder, G. R. (Ed.), *Maximum Entropy and Bayesian Methods* (pp. 1–42). New York: Springer.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*. New York: Springer.

Carlin, B. P. & Lewis, T. A. (2009). *Bayesian Methods for Data Analysis* (3rd Ed.). New York: CRC Press. Chapters 2-4.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd Ed.). New York: CRC Press. Chapters 7 and 10-12.

Gelman, A. & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. Discussion of “Bayesian model selection in social research,” by A. Raftery. In Marsden, P. V. (Ed.), *Sociological Methodology 1995* (pp. 165–173). New York: Blackwell.

Gelman et al. (2020). Bayesian workflow. https://arxiv.org/abs/2011.01808 (soon to be a book).

Hudson, T. E. & Landy, M. S. (2012). Measuring adaptation with a sinusoidal perturbation function. *Journal of Neuroscience Methods*, *208*, 48–58.

Kruschke, J. K. (2010a). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

Kruschke, J. K. (2010b). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1(5)*, 658–676.

Kruschke, J. K. (2012). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*, 573-603.

Lewandowsky, S. & Farrell, S. (2011). *Computational Modeling in Cognition* (Sections 5.2–5.5). Washington, DC: Sage.

Lu, Z.-L. & Dosher, B. A. (2014). *Visual Psychophysics: From Laboratory to Theory*. Cambridge, Mass.: MIT Press. Chapter 10.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms* (Ch. 28-30). Cambridge, UK: Cambridge Univ. Press. Sections 3.2-3.3 and Chapters 28-30.

Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Science*, *6*, 421–425.

Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. (2014). Bayesian model selection for group studies — Revisited. *Neuroimage*, *84*, 971-985.

Rouder, J. N. & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682-689.

Sivia, D. S. & Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial* (2nd Ed.). Oxford, UK: Oxford University Press. Chapters 4 and 9.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J. & Friston, K.J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*, 1004–1017.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D. & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

I. Why use Bayesian methods for inference? A cautionary example.

Bem: presents a slew of experiments with small, but $p < 0.5$ effects consisting with rejecting the null hypothesis of no effect in 9 “time-reversed” experiments (precognition, retroactive priming, etc.)

Wagenmakers et al.: Several problems with interpreting these p-values: (1) Exploratory vs. confirmatory studies. (2) $P(D\,|\,H)$ is not the same a $P(H\,|\,D)$, i.e., with healthy skepticism about ESP, these results aren't convincing. (3) Null-hypothesis tests only estimate evidence against H_0, not evidence for H_1. A default Bayesian t-test shows weak to no evidence for H_0 over H_1 in these experiments.

Rouder: Finds fault with the separate tests made by Wagenmakers

Lindley's paradox: If the null is true, the distribution of p-values is flat. If a particular, small effect-size alternative is true, the distribution of p-values is tight around small values. For modest sample sizes a *p*-value of 0.04 may be a bit more likely under H_1 than under H_0 although nowhere near as more likely as that p-value might make you believe. But, with a much larger sample size, that p-value will be evidence *for* H_0.

Suggests a meta-analysis that combines the studies for a single Bayes factor and results in moderately strong evidence (49:1) for H_1. This still is strongly outweighed by a sensible prior, but less so than Wagenmakers suggested.

II. Sampling parameters from the posterior

Why do we need sampling?

1. Want to compute a confidence interval for a parameter from a posterior.
2. Want to do model comparison and need to integrate over the posterior. This can be computationally infeasible, so summing over a sample can be a much easier approximation. For example, the Bayesian calculation of model probability requires such an integral: $p(M_i\,|\,y) \propto p(M_i)p(y\,|\,M_i)$ and the latter likelihood requires such an integration:

$$p(y\,|\,M_i) = \int p(y\,|\,\theta, M_i)p(\theta\,|\,M_i)d\theta.$$

Inverse cdf method

Rejection sampling: Sample from easy-to-sample $Mg(x)$ and accept the sample if a uniform sample (uniform from 0 to 1) is less than $f(x)/Mg(x)$.

Importance sampling (for calculating $E\left(f(x)\right)$): Sample from $q(x)$ but weight samples by $\frac{P^*(x)}{Q^*(x)}$, so you don't need to figure out how to sample from *p* nor how to normalize *p* or *q*.

MCMC: Markov Chain Monte Carlo methods

Gibbs sampling: iteratively draw a new value of θ_i conditional on current values of $\left\{\theta_{j\neq i}\right\}$ for a fixed order of visiting different values of i.

Metropolis algorithm:

(1) Draw initial parameter set θ_0 for which $p(\theta_0 \mid y) > 0$ from rough approximate distribution $p_0(\theta)$

(2) For $t = 1,2,\cdots$

a. Sample a proposal θ^* from a symmetric "jump" distribution $J_t(\theta^* \mid \theta^{t-1})$

b. Set $r = \frac{p(\theta^* \mid y)}{p(\theta^{t-1} \mid y)}$, note that unnormalized posteriors suffice for this step

c. Set $\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r,1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$

Metropolis-Hastings algorithm corrects for asymmetric jump distribution

Sampling datasets from the posterior: Sample θ as above, then sample from $p(y \mid \theta)$

Model checking: compare predicted *y*'s to data

III. Bayesian model comparison, Bayes factor, and Occam factor

Compare model posterior probabilities:

$\frac{p(M_1 \mid y)}{p(M_2 \mid y)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(y \mid M_1)}{p(y \mid M_2)}$, i.e., the posterior odds are the prior odds times the likelihood ratio of the models. The latter term is called the Bayes factor. Note that each term (e.g., $p(y \mid M_1)$) is the normalizing term that we chose to ignore when estimating model parameters (Bayesian parameter estimation; previous lecture). We refer to these terms as the *evidence* for each model, and their ratio is the Bayes factor.

To compute evidence, suppose we consider a single-parameter model and write out the evidence term as before:

$p(y \mid M_1) = \int p(y \mid \theta, M_1)p(\theta \mid M_1)d\theta$. This integral computes the area under the curve (as a function of the parameter θ) $p(y \mid \theta, M_1)p(\theta \mid M_1)$. Recall that the posterior $p(\theta \mid y, M_1) \propto p(y \mid \theta, M_1)p(\theta \mid M_1)$. Often, this posterior is tightly concentrated around the MAP estimate $\hat{\theta}$. Thus, the integral is of a curve that is the curve of the prior $p(\theta \mid M_1)$ shrunk by the likelihood term $p(y \mid \theta, M_1)$ so that it now peaks at the MAP estimate. This area can then be approximated by the height of the integrand at $\hat{\theta}$ times the width, i.e., $\int p(y \mid \theta, M_1) p(\theta \mid M_1) d\theta \approx p(y \mid \hat{\theta}, M_1) p(\hat{\theta} \mid M_1) \sigma_{\theta|y}$. In this approximation, the first term is the likelihood of the MAP estimate. This likelihood is reduced by the product of the next two terms, the *Occam factor*. Now, suppose the prior $p(\theta \mid M_1)$ was flat over a region with width σ_θ. In this case $p(\hat{\theta} \mid M_1) = 1/\sigma_\theta$, so that the Occam factor becomes $\sigma_{\theta|y}/\sigma_\theta$, i.e., it is the degree to which the effective parameter space shrank when the data arrived, thus penalizing evidence for models with too large a parameter space. This calculation will penalize models with large numbers of parameters, and those with larger effective ranges (widths of the prior) of those parameters.

In the multiple-parameter case, we can approximate the log-likelihood function as a quadratic by measuring the Hessian matrix (the matrix of second derivatives $H = \dfrac{\partial^2 \log p(\theta \mid y, M_1)}{\partial\theta_i \partial\theta_j}$. This is the multiple-parameter generalization of the curvature we measured to derive Fisher information last time. It just measures how the log-likelihood curves. If you then approximate the entire log-likelihood function based on this quadratic, you are effectively saying the likelihood function itself is Gaussian:
$p(\theta \mid y, M_1) \approx p(\hat{\theta} \mid y, M_1) \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T H (\theta - \hat{\theta})\right)$ with covariance matrix H^{-1}. The Occam factor becomes $p(\hat{\theta} \mid M_1)\sqrt{\dfrac{(2\pi)^K}{|H|}}$, where K is the number of parameters, which is the "volume" under the exponential above.

A "non-Bayesian" alternative: cross-validation and the notion of overfitting

Leave-one-out cross validation

Symptom of overfitting: error of prediction begins to increase with more parameters (i.e., fitting noise).

The applicability of this method to binomial data seems poor

Example: Hudson/Landy

Approximations and other ad hoc model-comparison methods

Nested models and the nested-hypothesis test

$(2 \log \dfrac{p(y \mid \hat{\theta}_{\text{complex}}, M_{\text{complex}})}{p(y \mid \hat{\theta}_{\text{simple}}, M_{\text{simple}})} \sim \chi^2(K)$, where K is the number of additional parameters in the complex model. Problem: only useful for rejecting the simple model, but does not tell you when the simple model is better, so not useful for model comparison

Gelman suggests the DIC (Deviance Information Criterion):

Deviance is simply a measure of fit: $D_\theta(y) = -2\log p(y\,|\,\theta, M)$. Before, we compared deviance of a psychometric function to that of the saturated model. Here we use deviance to compare models.

DIC $= 2\hat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y)$, where $\hat{D}_{\text{avg}}(y)$ is the average deviance of the data averaged over draws of θ from the posterior, and $\hat{D}_{\hat{\theta}}(y)$ is the deviance based on a point estimate (usually the posterior mean) of θ. Stated differently, DIC $= \hat{D}_{\text{avg}}(y) + \left(\hat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y)\right)$. The first term is the average deviance of the model. The second term is an estimate of the effective number of parameters of the model (effective in the sense of taking into account how much of a constraint on θ is imposed by the prior). Models may be compared by difference in DIC values.

Akaike's Information Criterion (AIC)

Want to rate model by Kullback-Leibler (KL) divergence (distance) of model-predicted from true probabilities:

$$KL = \int p(y)\log\frac{p(y)}{p(y\,|\,\theta, M)}dy = \int p(y)\log p(y)dy - \int p(y)\log p(y\,|\,\theta, M)dy$$

First term is independent of model and parameters, so use 2nd term to do model comparison.

Second term is expected log likelihood. Measured log likelihood approaches its expectation with large amounts of data, so use that instead. KL distance is based on θ, but model fitting uses the same data to estimate $\hat{\theta}$, so measured log likelihood using $\hat{\theta}$ is a biased estimate. The AIC tries to correct for this.

AIC $= = 2\log p(y\,|\,\hat{\theta}, M) + 2K$ where K is the number of model parameters

Compare models by computing the difference in AIC values

For small sample sizes or large numbers of parameters, the corrected AIC is recommended: AICc $=$ AIC $+ \dfrac{2K(K+1)}{N-K-1}$, where N denotes the sample size.

Bayesian Information Criterion (BIC)

The BIC is an attempt to estimate the evidence for a model without integrating over possible values of θ based on a particular choice of prior distribution $p(\theta\,|\,M)$.

BIC $= -2\log p(y\,|\,\hat{\theta}, M) + K\log N$ where N is the number of datapoints on which the log likelihood is based.

As an estimate of log model evidence, one can use the BIC to compute an estimated Bayes factor:

$$\text{Bayes factor} \approx \exp\left(-\frac{1}{2}(BIC_1 - BIC_2)\right)$$

Group studies and the protected exceedance probability
Graphical models and hidden parameters. Difficulty of inference and estimation in such models.
Bayesian workflow
Software aids: WinBUGS, RBUGS, JAGS, MatJAGS, Stan/RStan

Homework (due 12/18, 2PM): Simulate a set of observers in a motion-adaptation task. You have four conditions: adaptation direction (adapt to leftward or rightward motion) combined factorially with attentional condition (attention on the adapter or diverted from the adapter). For each, you collect a psychometric function for left-right discrimination, without feedback, as a function of motion coherence of a brief test stimulus (where -1.0 means all the dots go to the left, 0.0 means the dots are moving in random directions, 1.0 means all the dots go to the right and, e.g., 0.5 means that half the dots go to the right and the other half move in random directions). Assume a cumulative normal psychometric function. You will compare models that allow for inter-subject differences (in effect size for adaptation aftereffect, i.e., change in PSE, and also in slope/sigma and possibly in left/right bias). (Note: a PSE here is the coherence value that leads to indifference as to whether the stimulus moves left or right.) You want to compare several models:

M1: There is no adaptation effect (i.e., the slopes in the four conditions for a subject are identical, and the PSEs in the four conditions for a given subject are identical)

M2: There is an effect on PSE from adaptation, but no attentional effect (thus, there are two PSEs per subject, shifted from each other in the appropriate way expected for a motion after-effect)

M3: There is also an effect of attention, enhancing the motion after-effect. Thus, there are four distinct PSEs per subject, in the order left-adapt-with-attention, left-adapt-without-attention right-adapt-without-attention right-adapt-with-attention

M4: There is also an effect of attention on slope, but you aren't sure what that effect is in advance. This is the same as model M3 except that you are allowing two values of slope per subject (with and without attention during adaptation).

So: in the grand scheme of things, simulate data from N subjects for one of the models, then do a Bayesian comparison of all models using Jeffreys priors (as constrained by each model) for slope and PSE. You can also do maximum-likelihood fits and compare models using AIC and/or BIC and compare those results to a true Bayesian model comparison. This is a huge assignment and I don’t expect anyone to do all of it, but see how far you get and try to learn a bit about practical Bayesian model comparison along the way.