Problem Sets Between Tests 1 and 2

*Only turn in problems that are **not** bracketed.* Bracketed problems are additional problems you can look at. Round brackets indicate problems that may help you with problems that are assigned; square brackets are additional problems on material that you should know, but you are not required to write up solutions; curly brackets are truly optional and may contain extra nuggets that you will not be required to know but may be interested in.

Additional assignments will be filled in over time.

notation	meaning
unbracketed	assigned problem – turn these in for grading
()	helper/warm-up problem
[]	additional problems (you are responsible for content, but don't turn them in)
{}	covers optional material

PS	Due	Source	Problems
9	Wed 3/8	Rethinking 6 Additional Problems	**6E3–6E4** entropy **1** entropy **2** entropy
10	Fri 3/17	Rethinking 6	**6M3** same data **6M4** narrow prior **6H1** compare models **6H4** train vs. test **6H5** train vs. test *Use the replacement code below in place of the R code 6.31 and 6.32.*
11	Wed 3/29	Rethinking 6	**6M2** selection v averaging **6M5–6M6** over/under fitting **6H2** plotting models **6H3** model averaging
12	Mon 4/2	Rethinking 7	TBA

Replacement Code

R Code 6.31

```
library(rethinking)
data(Howell1)
Howell <-
  Howell1 %>%  mutate(age.s = zscore(age))
set.seed(1000)      # so we all get the same "random" data sets
train <- sample(1:nrow(Howell), size = nrow(Howell) / 2)  # half of the rows
Howell.train <- Howell[ train, ]   # put half in training set
Howell.test  <- Howell[-train, ]   # the other half in test set
```

R Code 6.32

```
# You need to come up with mu and sigma
sum(dnorm(Howell.test$height, mu, sigma, log = TRUE))
```

Additional Problems

1 Making entropy larger.

a) Which is larger: $H(0.1, 0.3, 0.6)$ or $H(0.2, 0.2, 0.6)$?

b) Let $\boldsymbol{p} = \langle p_1, p_2, p_3 \rangle$ and let $\boldsymbol{q} = \langle p, p, p_3 \rangle$ where $p = \frac{p_1+p_2}{2}$. Compute $H(\boldsymbol{p})$ and $H(\boldsymbol{q})$. Which is larger?

c) Suppose a random process has n outcomes. Show that with one exception, there is always another random process that also has n outcomes, but has higher entropy? What is the one exception? (The exception is the random process with the maximal entropy among processes with n outcomes.)

Solution.

```
H <- function(p) - sum(p[p>0] * log(p[p>0]))
H(c(0.1, 0.3, 0.6))

## [1] 0.8979457

H(c(0.2, 0.2, 0.6))

## [1] 0.9502705
```

Let $h(p) = p\log(p) + (s-p)\log(s-p)$ where s is fixed. I've used little h because this is a little part of the full entropy where the probabilities of two events sum to s. I've left off the negative sign to simplify the derivative below.

$h'(p) = \log(p) + p\frac{1}{p} - \log(s-p) - (s-p)\frac{1}{s-p} = \log(p) - \log(s-p)$. So $h'(p) = 0$ when $p = s - p$. This means that the largest entropy happens when these two events have the same probability.

We play this game any time thre are two unequal probabilities, so the the maximal entropy is acheived when all the probabilities are equal.

(This can also be deonstrated by clever algebra and log rules, but I find this more instructive and less messy.)

2 Compute the entropy of tossing two coins two different ways:

a) Consider the outcomes to be 0, 1 or 2 heads.

b) Consider the outcomes to be HH, HT, TH, or TT.

How do the results compare? Can you generalize?