# EEEB UN3005/GR5005
# Homework - Week 07 - Due 02 Apr 2019

*Xun Zhao, xz2827*

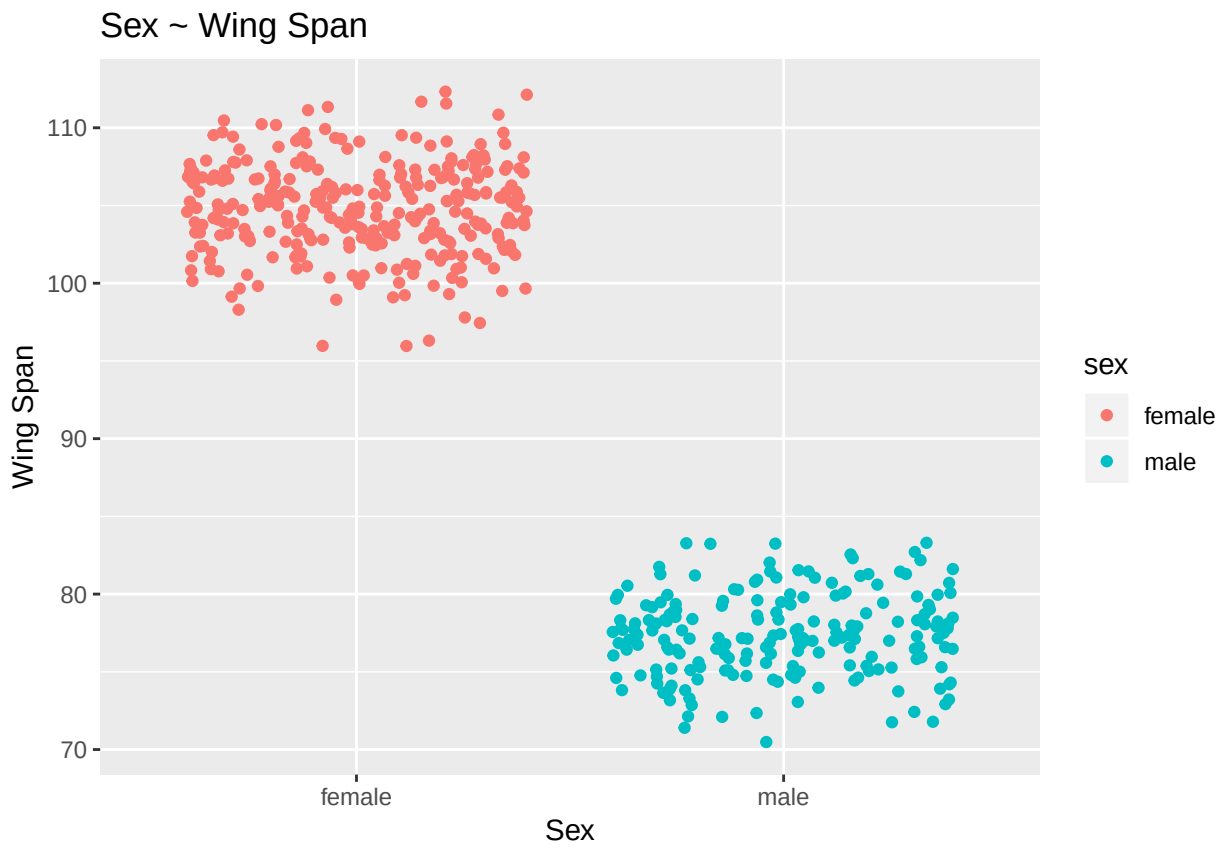
**Homework Instructions:** Complete this assignment by writing code in the code chunks provided. If required, provide written explanations below the relevant code chunks. Replace "USE YOUR NAME HERE" with your name in the document header. When complete, knit this document within RStudio to generate a pdf. Please review the resulting pdf to ensure that all content relevant for grading (i.e., code, code output, and written explanations) appears in the document. Rename your pdf document according to the following format: hw_week_07_firstname_lastname.pdf. Upload this final homework document to CourseWorks by 5 pm on the due date.

All of the following homework problems will use a dataset called `simulated_falcons.csv` which contains wingspan measures (in centimeters) for male and female falcons. For problems that ask you to create statistical models, use priors of `dnorm(0, 100)` for all intercept and coefficient parameters and a prior of `dcauchy(0, 10)` for all standard deviation parameters. Use explicit start values of 0 for all model parameters to ensure a good model fit. In addition, **you will have to use** `method = "SANN"` as an additional argument to `map()` to ensure all these models fit correctly. By default, `map()` uses `method = "BFGS"`, so we have to specify we want to use a different `method`. This bit of code will just be an extra part of your `map()` call in addition to your model code, `data`, and `start` arguments. Look at `?map` if you're unclear on the necessary code syntax.

## Problem 1 (2 points)

Import the `simulated_falcons` dataset, and create a jitter plot with the `sex` variable ("female" or "male") on the x-axis and the `wingspan` variable on the y-axis. This should give you a good idea as to the variation in falcon wingspan both between and within sexes.

```
d = read.csv('../Data/simulated_falcons.csv')
graph = ggplot(d, aes(x = sex, y = wingspan, color = sex)) +
    geom_jitter() +
    xlab('Sex') +
    ylab('Wing Span') +
    ggtitle('Sex ~ Wing Span')
plot(graph)
```




## Problem 2 (2 points)

Now fit a Gaussian model with wingspan as the outcome variable, and report the 99% PIs for the model parameters using `precis()`.

Based on these parameter estimates and what you visualized in Problem 1, do you think this model provides a good description of the data? Why or why not?

```
model.1 = map(
    alist(
        wingspan ~ dnorm(mu, sigma),
        mu ~ dnorm(0, 100),
        sigma ~ dcauchy(0, 10)),
    data = d,
    start = list(mu = 0, sigma = 0),
    method = 'SANN')
precis(model.1, prob = 0.99)
```

```
##         Mean StdDev  0.5% 99.5%
## mu     93.81   0.61 92.23 95.39
## sigma 13.71   0.43 12.59 14.82
```

**Answer:**

No, the model does not fit the true distribution of `wingspan`, because the variable is distributed as two parts, and there is no data points between these two parts. That is, the distribution is not Gaussian.

As the model's parameters indicate, it is very likely for `wingspan` to be around 94 (`mu`). However actually, there is no value between 85 and 95.

---

## Problem 3 (3 points)

Now fit a linear regression model with sex as a predictor of falcon wingspan. Note, since this variable is currently represented as a factor in the `simulated_falcons` dataset, you'll have to create a dummy variable for sex to use within your model. You can generate the dummy variable as you wish, coding either "male" or "female" with a value of 1.

Report the 99% PIs for the parameter estimates from this model. Based on these results, do you think this model provides a better description of the data? Why or why not?

```
d$sex.index = ifelse(d$sex == 'male', 1, 0)
model.2 = map(
    alist(
        wingspan ~ dnorm(mu, sigma),
        mu <- a + b * sex.index,
        a ~ dnorm(0, 100),
        b ~ dnorm(0, 100),
        sigma ~ dcauchy(0, 10)),
    data = d,
    start = list(a = 0, b = 0, sigma = 0),
    method = 'SANN')
precis(model.2, prob = 0.99)
```

```
##          Mean StdDev   0.5%  99.5%
## a      104.77   0.17 104.35 105.20
## b      -27.40   0.26 -28.07 -26.72
## sigma    2.86   0.09   2.63   3.09
```

---

**Answer:**

We can simply evaluate this model as follows:

When it is `male`, `sex.index = 1`.

$$\mu_{\text{male}} = a + b \times 1 \approx 77$$

When it is `female`, `sex.index = 0`.

$$\mu_{\text{female}} = a + b \times 0 \approx 104$$

This results fit the mean value based on the jitter graph, where `female` is around 105 and `male` is around 75. Thus, it is better than the first model, as the second model can describe two sexes' distribution differently.

---

## Problem 4 (3 points)

And now we come to a bit of a challenge. Can you fit a model that assumes falcon wingspan is a Gaussian outcome where both the mean wingspan AND the standard deviation in wingspan differs by sex? If you're able to fit the model for Problem 3, you have all of the conceptual tools needed to tackle this problem too. Think about how we go about modeling the Gaussian mean parameter in a typical linear regression model and apply those same strategies to modeling the standard deviation parameter...

Summarize the fit model parameters using 99% PIs. What do the parameter estimates indicate about the standard deviation in wingspan of male versus female falcons? Would you interpret this as a strong or weak effect?

```
model.3 = map(
    alist(
        wingspan ~ dnorm(mu, sigma),
        mu <- a.mu + b.mu * sex.index,
        sigma <- a.sigma + b.sigma * sex.index,
        a.mu ~ dnorm(0, 100),
        b.mu ~ dnorm(0, 100),
        a.sigma ~ dcauchy(0, 10),
        b.sigma ~ dnorm(0, 100)),
    data = d,
    start = list(a.mu = 0, b.mu = 0, a.sigma = 0, b.sigma = 0),
    method = 'SANN')
precis(model.3, prob = 0.99)
```

```
##           Mean StdDev   0.5%  99.5%
## a.mu    104.77   0.17 104.32 105.21
## b.mu    -27.40   0.26 -28.07 -26.74
## a.sigma   3.01   0.12   2.69   3.32
## b.sigma  -0.33   0.18  -0.81   0.14
```

---

**Answer:**

The mean values of two sexes are the same as the second model, `model.2`.

The `a.sigma` and `b.sigma` indicate that, for `male`, $\sigma_{\text{male}} \approx 2.7$, and for `female`, $\sigma_{\text{female}} \approx 3$. It fits the jitter graph, where `female`'s values are distributed in a wider range, while `male`'s range is smaller.

I am not sure which statistical test I should use to judge whether the variances are significantly different (Maybe it is *F-test*).

However, the difference between them is already 10% of $\sigma_{\text{female}}$, which is a large difference. So I would interpret this as a strong effect.

---