STA 360/602L: Module 3.3

The normal model: introduction and motivating examples

Dr. Olanrewaju Michael Akande

Motivating example: job training

- In the 1970s, researchers in the U.S. ran several randomized experiments intended to evaluate public policy programs.
- One of the most famous experiments is the National Supported Work (NSW) Demonstration, in which researchers wanted to assess whether or not job training for disadvantaged workers had an effect on their wages.
- Eligible workers were randomly assigned either to receive job training or not to receive job training.
- Candidates eligible for the NSW were randomized into the program between March 1975 and July 1977.
- For more details, read Lalonde, R. J. (1986) and Dehejia, R., and Wahba, S. (1999).

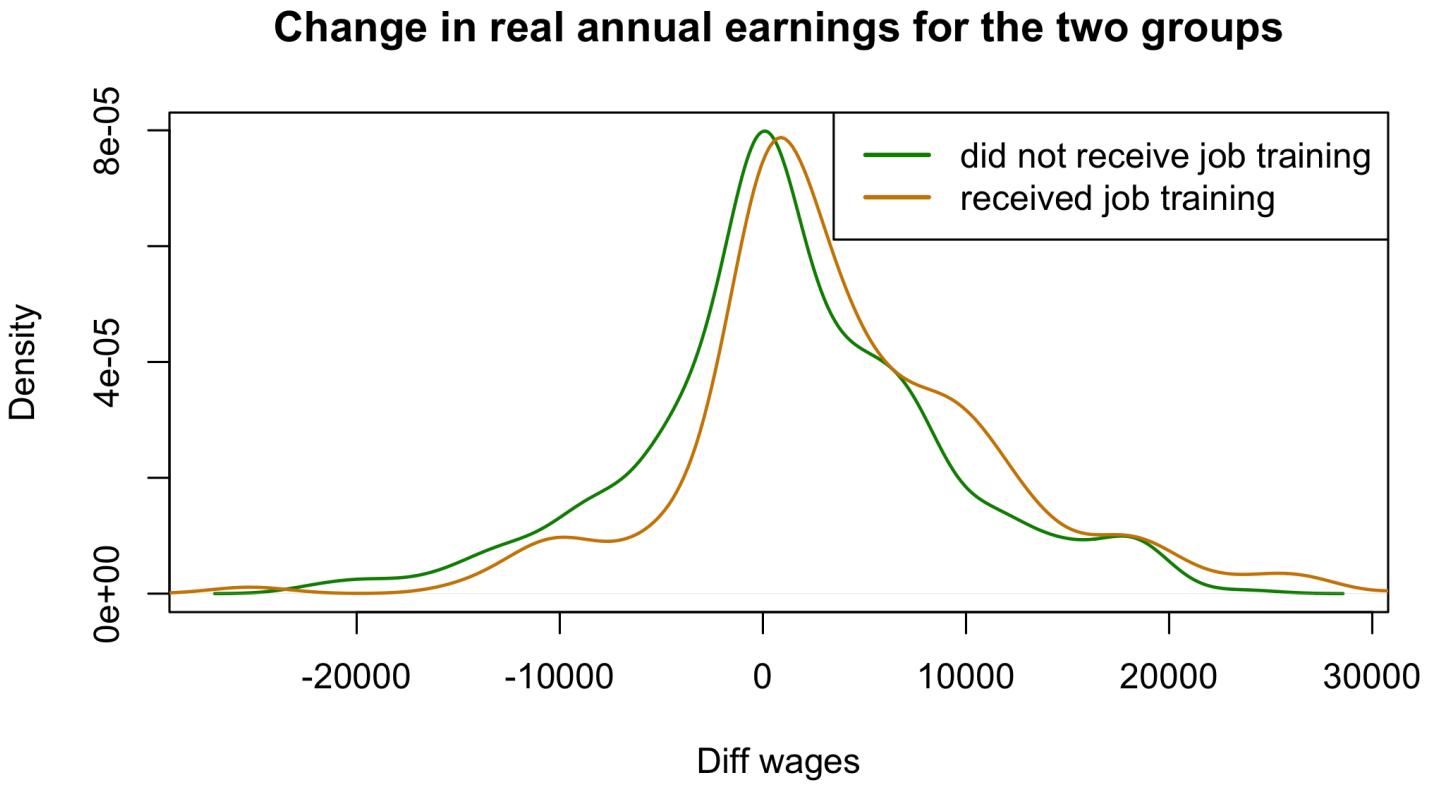
Motivating example: job training

- Setup:
  - **Pre-training wages**: real annual earnings in 1974 before training.
  - Two groups: some participants received job training and the rest did not.
  - **Post-training wages**: real annual earnings in 1978 upon completion of training.
- Question of interest: is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training?
- The original study really is a causal inference setup, but the data used in this example only uses a subset of the data.
- The data is richer than what we will use it for (i.e., there are covariates we can control for) but we will only focus on the pre and post wages for the two groups.

Job training: the data

- Data:
  - No training group (N): sample size $n_N = 429$.
  - Training group (T): sample size $n_A = 185$.
  - **Diff wages**: Post-training wages -- Pre-training wages.
- Summary statistics for change in annual earnings:
  - $\bar{y}_N = 1364.93$; $s_N = 7460.05$
  - $\bar{y}_T = 4253.57$; $s_T = 8926.99$
- Wages/income are well known to be approximately normally distributed. Let's look at the distribution of "change in annual earnings" for the two groups.

JOB TRAINING: THE DATA



Not completely normal but not too far off either. Lots of overlap between the two groups.

Model for changes in earnings

- $$y_i^{(T)} \sim \mathcal{N}(\mu_T, \sigma_T^2)$$

  $$y_i^{(N)} \sim \mathcal{N}(\mu_N, \sigma_N^2)$$

- Want posterior distribution of $\mu_T - \mu_N$. Specifically, we would like to compute $\Pr[\mu_T > \mu_N | Y_T, Y_N)$ or equivalently, $\Pr[\mu_T - \mu_N > 0 | Y_T, Y_N)$.
- Inference for $\mu_T - \mu_N$ can be complicated in frequentist paradigm when $\sigma_T^2 \neq \sigma_N^2$.
- Use approximate t-distributions based on the Welch-Satterthwaite degrees of freedom.
- Trivial with Bayesian inference
- By the way, also trivial to compute $\Pr[\sigma_T^2 > \sigma_N^2 | Y_T, Y_N)$ with Bayesian inference, which we will do later.
- How to do posterior inference for such normal models?

Another example: Pygmalion study

- Pygmalion effect is a phenomenon where expectation affects performance.
- Question of interest: do teachers' expectations impact academic development of children?
- Setup:
  - Researchers gave IQ test to elementary school children.
  - Randomly picked six children & told teachers that the test predicts them to **have high potential for accelerated growth**.
  - They randomly picked six children and told teachers that the test predicts them to have **NO potential for growth**.
  - At end of school year, they gave IQ test again to all students.
  - They recorded the change in IQ scores of each student.

Another example: Pygmalion study

- Data:
  - Accelerated group (A): 20, 10, 19, 15, 9, 18.
  - No growth group (N): 3, 2, 6, 10, 11, 5.
- Summary statistics:
  - $\bar{y}_A = 15.2$; $s_A = 4.71$.
  - $\bar{y}_N = 6.2$; $s_N = 3.65$.
- IQ test scores are also well known to be approximately normally distributed.
- Can't really check this assumption with only $n = 6$ observations.

Model for changes in scores

- $$y_i^{(A)} \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

  $$y_i^{(N)} \sim \mathcal{N}(\mu_N, \sigma_N^2)$$

- Once again, we want posterior distribution of $\mu_A - \mu_N$.
- As before, we would like to compute $\Pr[\mu_A > \mu_N | Y_A, Y_N) \equiv \Pr[\mu_A - \mu_N > 0 | Y_A, Y_N)$.
- We would also like to compute $\Pr[\sigma_A^2 > \sigma_N^2 | Y_A, Y_N)$.
- To answer both questions, let's learn the Bayesian normal model.

Normal distribution

- A random variable Y has a normal distribution, written as $Y \sim \mathcal{N}(\mu, \sigma^2)$, if the pdf is

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}}; \quad y \in (-\infty,\infty),\ \ \mu \in (-\infty,\infty),\ \ \sigma \in (0,\infty).$$

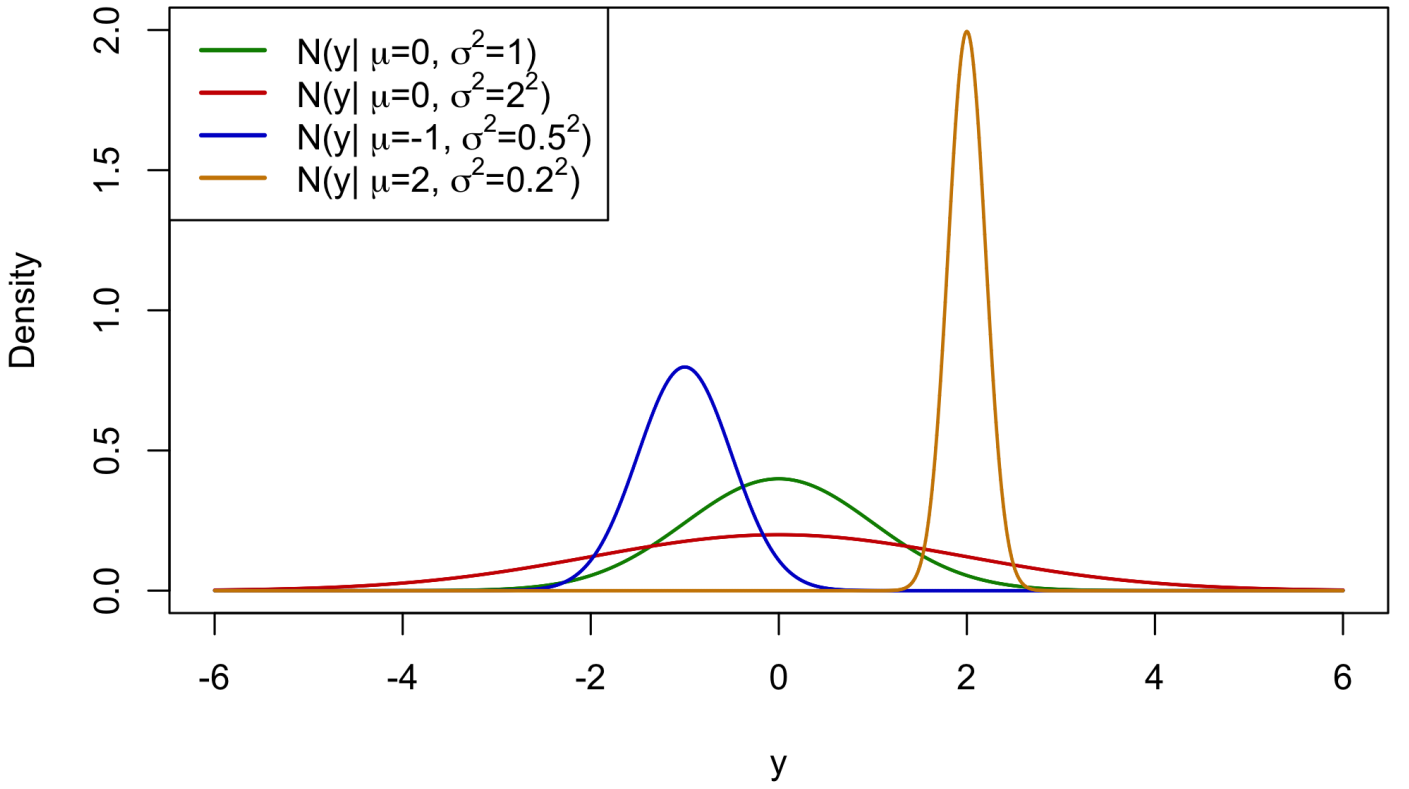
  where μ is the mean and σ^2 is the variance.

- It is also common (and would often be more convenient for our purposes) to write the pdf in terms of precision, τ, where $\tau = 1/\sigma^2$.

- In that case, the pdf is instead

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\tau^{\frac{1}{2}}\, e^{-\frac{1}{2}\tau(y-\mu)^2}; \quad y \in (-\infty,\infty),\ \ \mu \in (-\infty,\infty),\ \ \tau \in (0,\infty).$$

Example normal distributions



Comments on the normal distribution

- It is amazing how often real data are close to normally distributed.
- Likely a consequence of CLT -- sums and means of independent random variables tend to be approximately normally distributed.
- Occurs under very general conditions.
- Normality?
  - Height, weight and other body measurements,
  - Income\wages\earnings,
  - Cumulative hydrologic measures such as annual rainfall or monthly river discharge,
  - Errors in astronomical or physical observations,
  - Many more examples!

Properties of the normal distribution

- Mean, median and mode are all the same (μ).
- Symmetric about the mean μ.
- 95% of the density (95% probability) within $\pm 1.96\sigma$ (approximately two standard deviations) of the mean.
- If $X \sim \mathcal{N}(\theta, s^2)$ and $Y \sim \mathcal{N}(\mu, \sigma^2)$ with $X \perp Y$, then

$$aX + bY \sim \mathcal{N}(a\theta + b\mu, a^2 s^2 + b^2 \sigma^2),$$

  for constants a and b.
- When independence does not hold, the sum of two normally distributed random variables is still normally distributed.
- However, when that is the case, we must account for the correlation in the variance term.

Notes on normal distribution in R

- `rnorm`, `dnorm`, `pnorm`, `qnorm` in R take mean and **standard deviation** σ as arguments.
- If you use the variance σ^2 instead you will get wrong answers!
- For example, `rnorm(n,m,s)` generates n normal random variables with mean m and standard deviation s, that is, $\mathcal{N}(m, s^2)$.

What's next?

Move on to the readings for the next module!