Stats 401 Lab 8

Sanjana

10/25/2018

Outline

- Quick Reminder: If you are thinking about withdrawing from the course, the deadline is **November 9th**!
- Review of expectation, variance and covariance
- Midterm Q3 solved
- Introduction to confidence intervals

Review: Univariate Random Variable

- Recall: A random variable X is a value whose outcome is determined by a random process.
  Can be be thought of as a random number with probabilities assigned to its outcomes.
- Each random variable X has a mean (μ_x) and variance (σ_x^2) associated with it. Then
  - $\mu_x = \mathrm{E}[X]$.
  - $$\begin{aligned}
\sigma_x^2 = \mathrm{Var}(\mathbf{X}) &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\
&= \mathrm{E}[X^2 + (\mathrm{E}[X])^2 - 2X\mathrm{E}[X]] \\
&= \mathrm{E}[X^2] + \mathrm{E}[(\mathrm{E}[X])^2] - \mathrm{E}[2X\mathrm{E}[X]] \\
&= \mathrm{E}[X^2] + (\mathrm{E}[X])^2 - 2\mathrm{E}[X]\mathrm{E}[X] \\
&= \mathrm{E}[X^2] + (\mathrm{E}[X])^2 - 2(\mathrm{E}[X])^2 \\
&= \mathrm{E}[\mathbf{X^2}] - (\mathrm{E}[\mathbf{X}])^{\mathbf{2}}
\end{aligned}$$

Review: Univariate Random Variable

- Linear combinations of X: consider $Y = aX + b$
  - $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$
  - $\mathrm{Var}(aX + b) = a^2 \, \mathrm{Var}(X)$
- Normal Variable: let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$,
  Let $Y = aX + b$
  Then $\mathrm{E}[Y] = a\mu_x + b$ and $\mathrm{Var}(Y) = a^2\sigma_x^2$
  $Y \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$

Review: Bivariate Random Variables

- A bivariate random variable (X, Y) is a vector of length 2 whose values are each random variables.
- *Correlation* is a measure of the linear association between two random variables
  - Correlation is always between -1 and 1 (inclusive)
  - Correlation is symmetric: $\text{Cor}(X, Y) = \text{Cor}(Y, X)$
- *Covariance* is the unscaled version of correlation

$$\begin{aligned}
\mathbf{Cov(X, Y)} &= \mathbf{E}\left[(\mathbf{X} - \mathbf{E(X)})(\mathbf{Y} - \mathbf{E(Y)})\right] \\
&= \text{E}\left[(XY - X\text{E}(Y) - \text{E}(X)Y + \text{E}(X)\text{E}(Y)\right] \\
&= \text{E}[XY] - \text{E}[X\text{E}(Y)] - \text{E}[\text{E}(X)Y] + \text{E}[\text{E}(X)\text{E}(Y)] \\
&= \text{E}[XY] - \text{E}(X)\text{E}(Y) - \text{E}(X)\text{E}(Y) + \text{E}(X)\text{E}(Y) \\
&= \mathbf{E[XY] - E(X)E(Y)}
\end{aligned}$$

$$\mathbf{Corr(X, Y)} = \frac{\mathbf{Cov(X,Y)}}{\sqrt{\mathbf{Var(X)}}\sqrt{\mathbf{Var(Y)}}}$$

Review: Bivariate Random Variables

- $\mathrm{E}[(X, Y)] = (\mathrm{E}(X), \mathrm{E}(Y))$
- $\mathrm{Var}(X, Y) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix}$

- **Remember:** $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$

- Linear Combinations of multivariate $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$: consider $\mathbb{A}\mathbf{X}$
  - $\mathrm{E}(\mathbb{A}\mathbf{X}) = \mathbb{A}\mathrm{E}(\mathbf{X})$
  - $\mathrm{Var}(\mathbb{A}\mathbf{X}) = \mathbb{A}\mathrm{Var}(\mathbf{X})\mathbb{A}^T$

Review: Bivariate Normal Variables

If (X, Y) is bivariate normal where $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ then $aX + bY$ is also normal.

- $\mathrm{E}[aX + bY] = a\mu_X + b\mu_Y$
-

$$
\begin{aligned}
\mathbf{Var}(\mathbf{aX} + \mathbf{bY}) &= \mathbf{Cov}(\mathbf{aX} + \mathbf{bY}, \mathbf{aX} + \mathbf{bY}) \\
&= \mathrm{Cov}(aX + bY, aX) + \mathrm{Cov}(aX + bY, bY) \\
&= a\mathrm{Cov}(aX + bY, X) + b\mathrm{Cov}(aX + bY, Y) \\
&= a\mathrm{Cov}(aX, X) + a\mathrm{Cov}(bY, X) + b\mathrm{Cov}(aX, Y) + b\mathrm{Cov}(bY, Y) \\
&= a^2\mathrm{Cov}(X, X) + ab\mathrm{Cov}(Y, X) + ba\mathrm{Cov}(X, Y) + b^2\mathrm{Cov}(Y, Y) \\
&= a^2\mathrm{Var}(X) + ab\mathrm{Cov}(X, Y) + ab\mathrm{Cov}(X, Y) + b^2\mathrm{Var}(Y) \\
&= \mathbf{a^2 Var(X)} + \mathbf{2abCov(X, Y)} + \mathbf{b^2 Var(Y)} \\
&= a^2\sigma_X^2 + 2ab\mathrm{Cov}(X, Y) + b^2\sigma_Y^2
\end{aligned}
$$

So, if (X, Y) is bivariate normal as above, then
$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\mathrm{Cov}(X, Y))$

In-lab Activity 1: Midterm Q3

Qtn) X and Y are bivariate random variables with respective means $\mu_X = \mu_Y = 0$, standard deviations $\sigma_X = 1$ and $\sigma_Y = 2$ and correlation $\text{cor}(X, Y) = 0.5$. Find the distributions of $X + Y$ and $X - Y$.

Soln) $X + Y$ and $X - Y$ are both normally distributed since they are linear combinations of normal random variables. So, we need to find the following

- $\text{E}(X + Y)$ and $\text{E}(X - Y)$
- $\text{Var}(X + Y)$ and $\text{Var}(X - Y)$

We are given that
- $\text{E}(X) = 0$, $\text{Var}(X) = \sigma_X^2 = 1$
- $\text{E}(Y) = 0$, $\text{Var}(Y) = \sigma_Y^2 = 2^2 = 4$
- $\text{Cor}(X, Y) = 0.5$

So, $\text{Cov}(X, Y) = \text{Cor}(X, Y)\sqrt{\sigma_X^2 \sigma_Y^2} = \text{Cor}(X, Y)\sigma_X \sigma_Y = 0.5(1)(2) = 1$

In-lab Activity 1: Midterm Q3

- $E[X+Y] = E[X] + E[Y] = 0 + 0 = 0$ and
  $E[X-Y] = E[X] - E[Y] = 0 - 0 = 0$
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y) = 1 + 4 + 2(1) = 7$
- $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(-Y) + 2\text{Cov}(X,-Y) = 1 + 4 - 2(1) = 3$

So, $X+Y \sim \mathcal{N}(0,7)$ and $X-Y \sim \mathcal{N}(0,3)$

In-lab Activity 2

Let (W, X, Y) be a multivariate normal vector such that
$\mathrm{E}(X) = 2\mathrm{E}(Y) = 2$ and $\mathrm{E}(W) = 0$.
$\mathrm{Var}(X) = \mathrm{Var}(Y) = \mathrm{Var}(W) = 2$.
$\mathrm{Cor}(X, Y) = -0.5$, $\mathrm{Cor}(Y, W) = -0.5$,$\mathrm{Cor}(X, W) = 0$.

Find the distribution of $W + X - Y$.

In-lab Acitivity 2: Solution

$X - Y + W$ is a normal variable, since linear combinations of normal variables are normal.

- $\mathrm{E}(X - Y + W) = \mathrm{E}(X) + \mathrm{E}(Y) + \mathrm{E}(W) = 2 - 1 + 0 = 1$
- $\mathrm{Var}(\begin{bmatrix} X \\ Y \\ W \end{bmatrix}) = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) & \mathrm{Cov}(X, W) \\ \mathrm{Cov}(Y, X) & \mathrm{Var}(Y) & \mathrm{Cov}(Y, W) \\ \mathrm{Cov}(W, X) & \mathrm{Cov}(W, Y) & \mathrm{Var}(W) \end{bmatrix}$
  - $\mathrm{Cov}(X, Y) = \mathrm{Cor}(X, Y)\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} = -0.5\sqrt{22} = -1$
  - $\mathrm{Cov}(X, W) = \mathrm{Cor}(X, W)\sqrt{\mathrm{Var}(X)\mathrm{Var}(W)} = 0\sqrt{22} = 0$
  - $\mathrm{Cov}(Y, W) = \mathrm{Cor}(Y, W)\sqrt{\mathrm{Var}(Y)\mathrm{Var}(W)} = -0.5\sqrt{22} = -1$

So, $\mathrm{Var}(\begin{bmatrix} X \\ Y \\ W \end{bmatrix}) = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$

In-lab Acitivity 2: Solution

Then, $X - Y + W = \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ W \end{bmatrix} = \mathbb{A}\mathbf{X}$, where $\mathbf{X} = \begin{bmatrix} X \\ Y \\ W \end{bmatrix}$

$$\begin{aligned}
\text{Var}(X - Y + W) = \text{Var}(\mathbb{A}\mathbf{X}) &= \mathbb{A}\text{Var}(\mathbf{X})\mathbb{A}^T \\
&= \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2(1) + (-1)(-1) + 0(1) \\ (-1)1 + 2(-1) + (-1)1 \\ 0(1) + (-1)(-1) + 2(1) \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -4 \\ 3 \end{bmatrix} \\
&= 1(3) + (-1)(-4) + 1(3) \\
&= 10
\end{aligned}$$

So, $\mathbf{X} - \mathbf{Y} + \mathbf{W} \sim \mathcal{N}(\mathbf{1}, \mathbf{10})$

Confidence Intervals

- We are we interested in studying confidence intervals?
  - CIs essentially perform a two-sided hypothesis test and provide you with a estimate the true population value
- There are several natural uses for confidence intervals in regression:
  - estimating population coefficients (β)
  - comparing means of different populations
  - predicting future values (prediction interval)
  - predicting mean future values (confidence interval)

Confidence Intervals: formulae

- Recall from Stats250 that a $100(1-\alpha)\%$ confidence interval for a value is given by
  - $x \pm z_{\frac{\alpha}{2}} s.e(x)$ (population s.d. is known) or
  - $x \pm t_{(\frac{\alpha}{2}, n-2)} s.e(x)$ (population s.d. is unknown)
- *Approximate Interval for Linear Model* An approximate $100(1-\alpha)$ CI for β_1 is

$$\mathbf{b_1} - \mathbf{z}_{\frac{\alpha}{2}} \mathrm{SE}(\mathbf{b_1})$$

In-lab Activity 3: Constructing CI in R

Construct a 90% CI for the association between GPA and ACT scores

```
# read-in dataset
gpa <- read.table("gpa.txt",header=T)
# fit model and print coefficients summary
fit <- lm(GPA~ACT, data=gpa)
fit_coef_smry <- summary(fit)$coefficients; fit_coef_smry
```

```
##               Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 1.55870218 0.138016669 11.29358 2.694524e-27
## ACT         0.05780049 0.005549788 10.41490 1.015955e-23
```

```
beta <- fit_coef_smry["ACT","Estimate"]
SE <- fit_coef_smry["ACT","Std. Error"]
z <- qnorm(1-0.1/2) #for a 90% CI using normal approximation
cat("CI = [", beta-z*SE, ",", beta+z*SE, "]")
```

```
## CI = [ 0.04867191 , 0.06692908 ]
```

Lab Ticket

- Let X and Y be two random variables such that $\text{Var}(X) = 4$ and $\text{Var}(Y) = 9$. Find the range for the $Cov(X, Y)$
- Let X, Y be bivariate normal such that they are independent. Further, $X \sim \mathcal{N}(1, 4)$ and $Y \sim \mathcal{N}(2, 9)$.
  Find the distribution of the following:
  - $2X + 1$
  - $X - 2Y$
- Would a 90% confidence interval be broader or a 95%? Justify your answer.