

Stat 111 Spring 2011 Week 3
Central Limit Theorem, Covariance, t distribution

Presentations

1. Prove **The Central Limit Theorem** in the case where X_1, \dots, X_n are i.i.d. random variables, all with mean 0, variance σ^2 , and MGF $M_x(t)$. Define $S_n = \sum X_i$ and $Z_n = S_n/(\sigma\sqrt{n})$.
 - a) Find the MGF for Z_n and show that it converges to the MGF for a standard Normal r.v. as $n \rightarrow \infty$. That is, show that Z_n converges in distribution to $N(0, 1)$.
 - b) The CLT implies that certain distributions will be approximately Normal for large values of certain parameters. Give a list of examples.
 - c) Describe the Normal Quantile Plot for assessing the normality of a set of data values. Use this to demonstrate some of the approximations in part b.

2. The *covariance* between random variables X and Y is defined as

$$\text{Cov}(X, Y) = \sigma_{xy} = E((X - \mu_x)(Y - \mu_y)) = E(XY) - E(X)E(Y).$$

- a) Show that, for constants a_1, a_2, b_1 and b_2 , $\text{Cov}(a_1 + b_1X, a_2 + b_2Y) = b_1b_2\text{Cov}(X, Y)$.
- b) Show that $\text{Var}(X + Y) = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$, and that, more generally,

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \left(\sum_{i=1}^n \text{Var}(X_i)\right) + 2\left(\sum_{i < j} \text{Cov}(X_i, X_j)\right).$$

What is $\text{Var}(X - Y)$?

- c) The *correlation* between X and Y is defined as $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$. Use the fact that $\text{Var}(X+Y)$ and $\text{Var}(X-Y)$ are both non-negative to prove that $-1 \leq \rho_{xy} \leq 1$.
- d) Show that if X and Y are independent, then $\text{Cov}(X, Y) = 0$.
- e) Suppose $Z \sim N(0, 1)$ and define $Y = SZ$, where S is -1 or 1 , each with probability 0.5 . Show that Y and Z are *uncorrelated* ($\rho_{sz} = 0$) but not independent.
- f) Suppose X_1, \dots, X_n are i.i.d. (not necessarily Normal) with mean μ_x and variance σ_x^2 . Find the covariance between \bar{X} and X_i , and between \bar{X} and $X_i - \bar{X}$.
- g) Derive the mean of $(X_i - \bar{X})^2$ and of $\sum_{i=1}^n (X_i - \bar{X})^2$. Show that the sample variance $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimate of σ_x^2 .

3. Prove that, for random variables $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, the sample average \bar{X} is independent of the sample variance s^2 .

a) Let Z_1 and Z_2 be independent standard Normal random variables. Define $\bar{Z} = (Z_1 + Z_2)/2$ and $D = Z_1 - \bar{Z}$. Show that \bar{Z} and D are uncorrelated.

b) Find the joint density of \bar{Z} and D and conclude that they are independent.

c) For $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, we can write

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim N_n(\mu \mathbf{1}, \sigma^2 \mathbf{I}_n)$$

to mean that X follows a n -dimensional Multivariate Normal distribution with $n \times 1$ mean vector $\mu \mathbf{1}$ (μ repeated n times) and $n \times n$ covariance matrix $\sigma^2 \mathbf{I}$ (i.i.d. observations implies the covariance matrix is proportional to the identity matrix). In general, if $Y \sim N_n(\boldsymbol{\mu}, \mathbf{V})$, the density function for the n -vector \mathbf{y} is

$$f_{\mathbf{y}}(\mathbf{y}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right].$$

Derive the density function for $Y = \mathbf{A}X$, where \mathbf{A} is an $n \times n$ invertible matrix. Section 14.4 of Rice gives some properties of vector random variables.

d) Define \mathbf{A} to be the matrix that transforms X into the vector Y , where

$$Y = \mathbf{A}X = \begin{pmatrix} \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}$$

Find the covariance matrix for Y , and explain how it implies the independence result.

e) Argue that this result implies that \bar{X} is independent of s^2 , the sample variance of the X_i 's.

4. Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu_x, \sigma_x^2)$.

- What is the distribution of $\sum_{i=1}^n (X_i - \mu_x)^2$?
- What are the distributions of $(\bar{X} - \mu_x)^2$ and of $n(\bar{X} - \mu_x)^2$?
- Derive the distribution of $S = \sum_{i=1}^n (X_i - \bar{X})^2$. You'll need the result from presentation 3.
- Show that the sample variance $s_x^2 = S/(n-1) \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2\sigma_x^2}\right)$, meaning that $\frac{(n-1)s_x^2}{\sigma_x^2} \sim \chi_{(n-1)}^2$.
- The *Student's t* distribution is defined as follows: if $Z \sim N(0, 1)$ is independent of $X \sim \chi_m^2$, then

$$T = \frac{Z}{\sqrt{X/m}} \sim t_{(m)}, \text{ and } f_t(t) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\Gamma(m/2)} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2}.$$

Let $T = \frac{\bar{X} - \mu_x}{s_x/\sqrt{n}}$, for $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu_x, \sigma_x^2)$. Show that $T | \mu, \sigma \sim t_{(n-1)}$.

- A common *noninformative prior* distribution for μ_x and σ_x^2 is $p(\mu_x, \sigma_x^2) \propto 1/\sigma_x^2$, for $\sigma^2 > 0$ (an improper Uniform prior on μ_x and $\log(\sigma_x^2)$). For $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu_x, \sigma_x^2)$ and $T = \frac{\mu_x - \bar{x}}{s_x/\sqrt{n}}$, show that $T | \bar{x}, s_x \sim t_{(n-1)}$. That is, show that this prior specification implies the same distribution for T whether you condition on the data (as in part e) or on the unknown parameters.

Problems to turn in:

- Suppose $X_1 \sim \text{Gamma}(\alpha_1, \lambda)$, independent of $X_2 \sim \text{Gamma}(\alpha_2, \lambda)$.

$$\text{Define } Y_1 = \frac{X_1}{X_1 + X_2}, \text{ and } Y_2 = X_1 + X_2.$$

- Show that $Y_1 \sim \text{Beta}(\alpha_1, \alpha_2)$, independent of $Y_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda)$. Use the general formula for a bivariate change of variables:

If $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$, with $X_1 = h_1(Y_1, Y_2)$ and $X_2 = h_2(Y_1, Y_2)$, then

$$f_{y_1 y_2}(y_1, y_2) = f_{x_1 x_2}(h_1(y_1, y_2), h_2(y_1, y_2)) |\det[\mathbf{J}]|, \text{ where } \mathbf{J} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix}.$$

b) If $R = \frac{X_1/\alpha_1}{X_2/\alpha_2}$, then R follows an $F(2\alpha_1, 2\alpha_2)$ distribution (or if $X_1 \sim \chi_{(n_1)}^2$ independent of $X_2 \sim \chi_{(n_2)}^2$, then $\frac{X_1/n_1}{X_2/n_2} \sim F(n_1, n_2)$). Find the mean of the $F(2\alpha_1, 2\alpha_2)$ distribution by finding the mean of X_1 and of $1/X_2$, and by using the fact that X_1 and X_2 are independent.

2. Use the Central Limit Theorem (CLT) to approximate probabilities about 6-sided dice throws (outcomes are $1, 2, \dots, 6$, each with probability $1/6$). You can use `pnorm` in R or a Normal table to get standard Normal probabilities (typing `pnorm(z)` in R returns the probability that a standard Normal r.v. is less than z).

a) Simulate the distribution of the sum of the rolls when you throw 20 dice. There are more efficient ways to do this in R, but to give you experience with a loop, type (you don't have to type the lines that begin with `#`; these are comments):

```
y = rep(0,1000)
# create a vector of 1000 0's that you will load with your random draws.
for(i in 1:1000){
# let the counter i range from 1,2,...,1000
  x = sample(c(1,2,3,4,5,6), 20, replace=T)
# draw a sample of size 20, with replacement, from the values 1,...,6, with
# equal probabilities for each value.
  y[i] = sum(x)
# load the sum of the x vector into the ith element of y.
}
```

Now `y` contains 1000 simulated sums of 20 dice rolls. Make a pdf of a histogram and normal quantile plot and print these out. Does the distribution appear approximately Normal?

b) Find the approximate theoretical 0.9 percentile of the distribution of the sum of 20 dice rolls. Compare your results to the simulated values. Type `quantile(y, .9)` to find the sample 90th percentile of your 1000 simulated values.

c) Find the approximate theoretical probability that the sum of 20 dice rolls will exceed the value 80. Use the Normal approximation with the *continuity correction* to adjust for the discreteness of the sum. Compare your answer to the simulation. Type `mean(y>80)` to see the proportion of your values over 80.

d) What are the 5% most extreme values for the *average* of $n = 20$ independent rolls of a fair die? That is, find the set of values that would lead you to reject that the mean roll is $\mu = 3.5$, working at the $\alpha = 0.05$ significance level.