MAT 5030
Chapter 3:
Probability and Distributions

Kazuhiko Shinki

Wayne State University

Random Sampling

Random sampling

To randomly choose numbers from a finite set, use the ”sample” function. This function is useful for ***resampling methods*** . Resampling methods evaluate variability of statistics, such as mean and standard deviation, by simulation.

Random Sampling

```
> sample(1:10, 5) # 5 random numbers from 1 to 10 without replacement
[1] 6 5 9 2 8

> sample(1:10) # a permutation  of {1,2,...,10}
 [1]  2  6  7  8  4  3  5  9 10  1

> sample(1:10,replace=T)
>  # 10 random numbers from 1 to 10 with replacement
 [1] 10  4 10  2  6  2  8  4  5  3

> sample(1:10, 11)
>  # choose 11 numbers from 1 to 10 without replacement (impossible)
Error in sample(1:10, 11) :
  cannot take a sample larger than the population when 'replace = FALSE'

> sample(1:10, 11, replace=T)
>   # choose 11 numbers from 1 to 10 with replacement
 [1] 10 10  5  3  8  7  3  6  2  6  7
```

Random Sampling

```
> sample(c(1,2,4,8), 3) # a sample from {1,2,4,8}
[1] 2 1 4

> sample(c("A","B","C"), 5, replace=T)
[1] "B" "C" "C" "C" "A"

> sample(c("A","B"), 10, replace=T, prob=c(0.2,0.8))
># probability of A is 0.2, B is 0.8.
 [1] "B" "B" "B" "A" "B" "B" "B" "B" "B" "B"
```

Random Sampling

Question:

We have data with 5 observations: 5.1, 4.8, 3.9, 5.3, 4.1.

1. Calculate mean and S.D. of the data.
2. Estimate the standard error of the mean theoretically.
3. Estimate the standard error of hte mean by a resampling method.

Random Sampling

Recall that, when $X_1, \cdots, X_n$ are given, the standard error (or standard deviation) of the mean $\bar{X}$ is given by:

$$SE_{\bar{X}} = \frac{S.D.\text{ of } X}{\sqrt{n}} \tag{1}$$

It implies that when n is large, the variability of the mean $\bar{X}$ becomes arbitrarily small. This coincides with accepted fact that the mean of many observations is stable.

Random Sampling

Sample Code:

```
# (1)
> X <- c(5.1, 4.8, 3.9, 5.3, 4.1)
> c(mean(X), sd(X))
[1] 4.640000 0.614817

# (2)
> sd(X)/sqrt(5)   # (standard deviation)/(sqrt of sample size)
[1] 0.2749545

# (3)
> M <- numeric(100)  # 100-dim vector
> for (i in 1:100){
+ M[i] <- mean(sample(X, replace=T))  # We simulate a sample mean 100 times
+ }
> sd(M) # standard deviation of the mean
[1] 0.2549166
```

Random Sampling

The method (3) is called ***bootstrapping*** . The estimate (0.2549166) tends to be slightly smaller than the answer in (2), but still a good estimate. The bias becomes negligible when the sample size becomes large.

Exercise 1:
Calculate the expected value for the standard deviation of $\boldsymbol{M}$ in the previous slide. (Hint: Try all $\mathbf{5^5 = 3125}$ permutations.)

Random Variables

Random variables

A ***random variable*** is a map from a probability space to a set of numbers (range).

Example: Flip a coin twice

- Probability Space $\Omega := \{(H, H), (H, T), (T, H), (T, T)\}$
- Let X be the total number of heads, then X is a random variable.

$$\begin{aligned} X : (H, H) &\mapsto 2 \\ X : (H, T) &\mapsto 1 \\ X : (T, H) &\mapsto 1 \\ X : (T, T) &\mapsto 0 \end{aligned}$$

Random Variables

Discrete random variables

A random variable is called ***discrete***, when the range is discrete (roughly speaking, the number of possible values are countable).

Probability mass function (discrete)

For a discrete random variable X, the probability of $X = k$ is written as $P(X = k)$ (or $p(k)$, $f(k)$) (k runs all possible values of X), and is called a ***probability mass (or point) function***.

Continuous random variables

A random variable is called ***continuous*** , when the range is continuous.

Density function (continuous)

$f(x)$ is called a ***density function*** of X if

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{2}$$

We can assume $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

Random Variables

Cumulative distribution function (CDF)

$F(x) = P(X \leq x)$ is called a ***cumulative distribution function (CDF)*** .
Note that if $F(x)$ is differentiable for the whole range of X,

$$F'(x) = f(x) \tag{3}$$

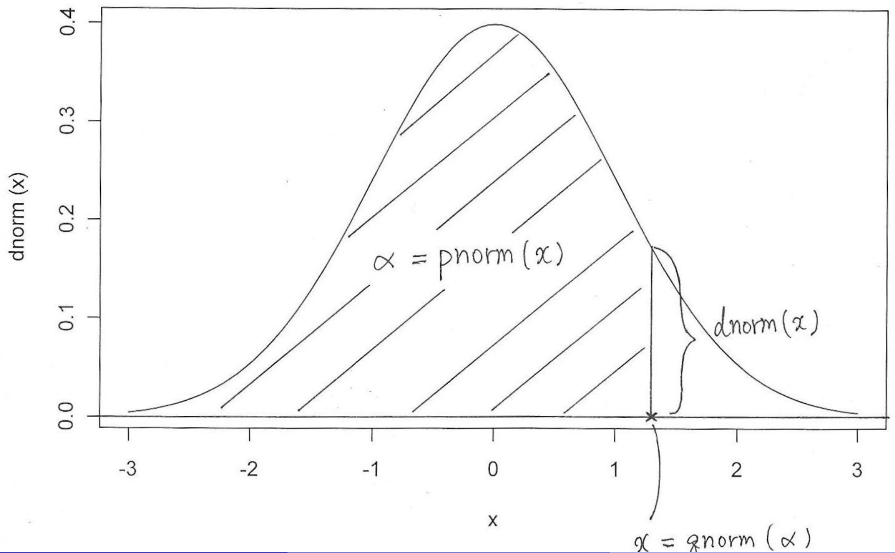
Random Variables

R Function for random variables

- d+'name' : mass or density function (f)
- p+'name' : cumulative distribution function (CDF:F)
- q+'name' : quantile function (inverse of CDF: F^{-1})
- r+'name' : random number (X)

Random Variables

R Function for random variables



Normal distribution:

Recall that the ***normal density function*** is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ is the expectation and $\sigma > 0$ is the standard deviation of the random variable.

We write $X \sim N(\mu, \sigma^2)$ when X has the above distribution.

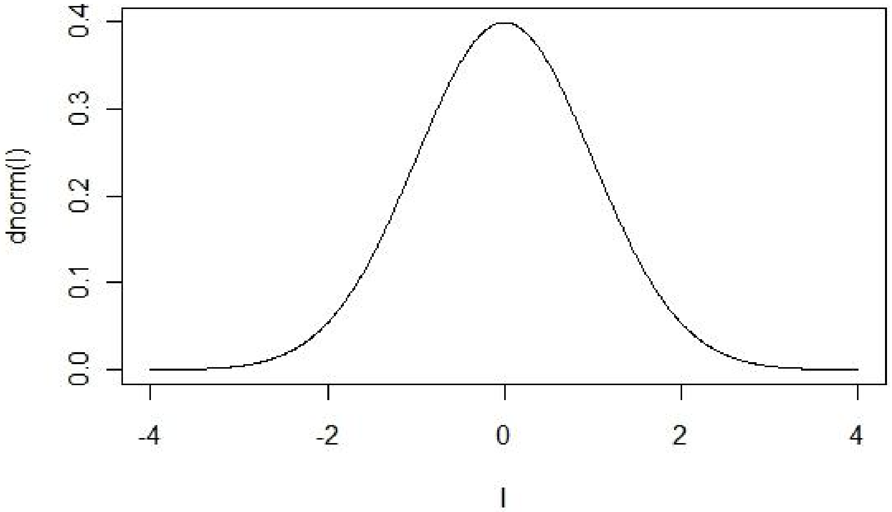
Example: standard normal

```
> dnorm(0) # f(0)
[1] 0.3989423
> dnorm(-2:2) # dnorm of -2, -1, 0, 1, 2
[1] 0.05399097 0.24197072 0.39894228 0.24197072 0.05399097

I <- 0.01*(-400:400) # -4, -3.99, ..., 3.99, 4
plot(I, dnorm(I),type="l")
# x: I, y:dnorm(I)
# type = line
```

Random Variables

Example: standard normal



Random Variables

Example: standard normal

```
> pnorm(0) # F(0)
[1] 0.5

> pnorm(1.96) # F(1.96)
[1] 0.9750021

> qnorm(0) # find x such that F(x) = 0
[1] -Inf

> qnorm(0.5) # find x such that F(x) = 0.5
[1] 0

> qnorm(0.975) # find x such that F(x) = 0.975
[1] 1.959964
```

Example: standard normal

```
> rnorm(10) # 10 standard normal random numbers
 [1] -0.67800199 -0.53466892 -0.64056387 -0.41621956  0.18128060
 [6] -0.59565417 0.09202977 -1.48117833  0.53581163 -1.80316248
```

Example: normal $N(0, 2^2)$

```
> dnorm(0, mean = 0, sd = 2) # dnorm(0) for N(0,4)
[1] 0.1994711

> pnorm(2*1.96, mean = 0, sd = 2) # pnorm(3.92) for N(0,4)
[1] 0.9750021

> qnorm(0.975, mean = 0, sd = 2) # qnorm(0.975) for N(0,4)
[1] 3.919928
```

See ‘help(rnorm)’ for more details.

Example: other random variables

Distribution		Parameter(s)	R functions (*1)
binomial	(discrete)	size, prob	binom
uniform	(continuous)	min, max	unif
normal	(continuous)	mean, sd	norm
χ^2	(continuous)	df	chisq
t	(continuous)	df	t
F	(continuous)	df1, df2	f
exponential	(continuous)	rate	exp
gamma	(continuous)	shape, scale	gamma

(*1) Add 'd", "p", "q" or "r" before the name of a distribution.

Random Variables

Examples:

```
> rbinom(8, size = 100, prob=0.7)
>   # Play 100 chess games each day.
>   # Your probability to win each game is 70%.
>   # What are the numbers of wins a day for 8 days?
[1] 60 72 79 75 68 77 71 66

> pbinom(65, size = 100, prob=0.7)
>    # What's the probability that you win 65 times or less?
[1] 0.1628583

> pt(2, df=4)
>   # what is the probability that a t_4 r.v. <= 2 ?
[1] 0.941941
```

Random Variables

χ^2 random variable:

The ***chi-square random variable*** χ^2_n with n degrees of freedom is generated by:

$$\chi^2_n = X_1^2 + X_2^2 + \cdots + X_n^2$$

where $X_1, X_2, \cdots, X_n$ are independent standard normal random variables.

Random Variables

Example: χ^2 random variable

Generate 1000 χ^2_3 random variables in two ways:

- "rchisq" function, and
- "rnorm" function,

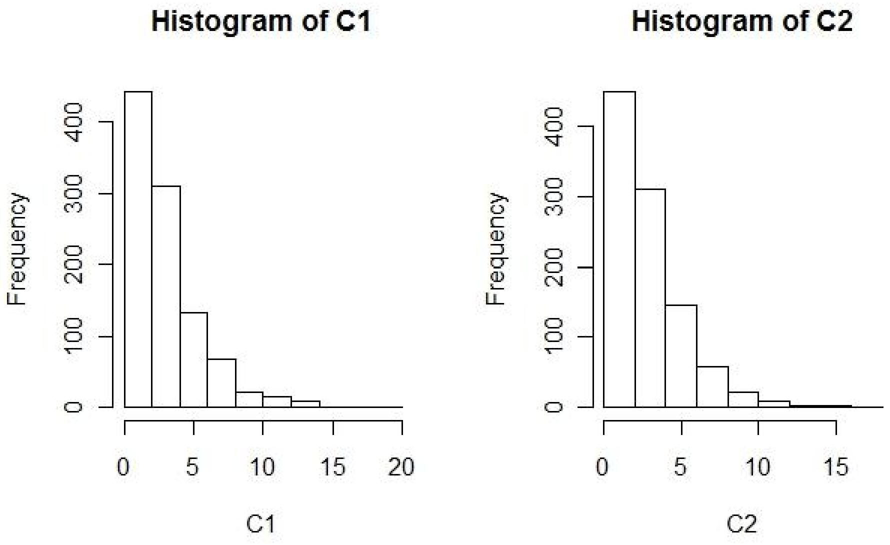
and compare the means and standard deviations.

Random Variables

```
> C1 <- rchisq(1000, df=3)
>
> X1 <- rnorm(1000)
> X2 <- rnorm(1000)
> X3 <- rnorm(1000)
> C2 <- X1^2 + X2^2 + X3^2
>
> rbind( c(mean(C1), sd(C1)), c(mean(C2),sd(C2)) )
         [,1]     [,2]
[1,] 2.947082 2.488845
[2,] 2.833078 2.348588

par(mfrow = c(1,2)) # align 2 graphs in one window
hist(C1)
hist(C2)
```

Random Variables



Random Variables

t random variable:

The ***t-random variable*** T_n with n degrees of freedom is generated by:

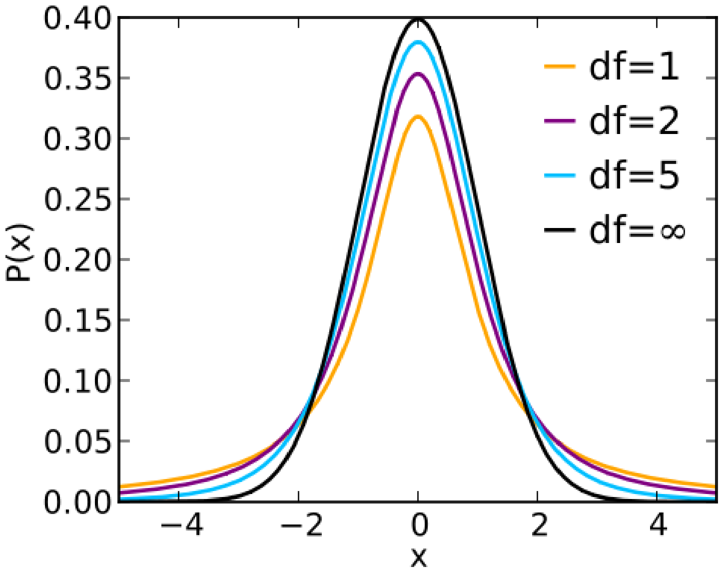
$$T_n = \frac{Z}{\sqrt{(X_1^2 + X_2^2 + \cdots + X_n^2)/n}}$$

where $Z, X_1, X_2, \cdots, X_n$ are independent standard normal random variables. Note that $(X_1^2 + X_2^2 + \cdots + X_n^2)$ is the same as χ_n^2.

Random Variables

The t distribution is symmetric and has heavier tails when n is smaller. When $n \to \infty$, t-distribution gets closer and closer to the standard normal distribution, since $(X_1^2 + X_2^2 + \cdots + X_n^2)/n$ in the denominator has an arbitrarily small variability when n gets large.

Random Variables



Exercise 2:

Generate 10,000 t-random variables with 4 degrees of freedom in two ways:

- by using the ”rt” function, and
- by using the ”rnorm” function,

and compare the means and standard deviations.

F random variable:

An ***F-random variable*** F with the numbers of degrees of freedom $df_1 = m$ and $df_2 = n$ is generated by:

$$F = \frac{\chi_m^2/m}{\chi_n^2/n} = \frac{(X_1^2 + X_2^2 + \cdots + X_m^2)/m}{(Y_1^2 + X_2^2 + \cdots + Y_n^2)/n}$$

where χ_m^2 and χ_n^2 are independent chi-square random variables with degrees of freedom m and n respectively, and $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are independent standard normal random variables.

Exercise 3:

Calculate the 90th percentile of the F(3,5) distribution in two ways:

- by ”qf” function,
- by ”rchisq” function and simulation.

Other relationships among random variables

A couple of insightful diagrams:

- Casella and Berger, “Statistical Inference,” *Duxbury Press*; 2 edition, pp.627.
- L.M. Leemis and J.T. Mcquestion, “Univariate Distribution Relationships,” *the American Statistician*, pp.45-53, 62(1), 2008.

Exercise 4:

(a) Generate 10,000 random numbers of t_5, standardize the numbers by its sample standard deviation, and calculate the proportion of observations which exceed 2.

(b) Repeat (a) for t_3 and t_{10}

(c) Calculate the theoretical probability that a t_5-random variable divided by its (theoretical) standard deviation exceeds 2, by using the quantile function `qt`.

- $\mathrm{Var}[T_n] = \frac{n}{n-2}$ where n is the number of degrees of freedom.