2. Probability and Sampling Distributions

Outlines

- ✓ R as a set of statistical tables
- ✓ Examining the distribution of a set of data
- ✓ Simulating the Sample Distribution of the Mean
- ✓ One and two sample tests

Probability and Sampling distribution

R as a set of statistical tables

- The **R** suite of programs provides a simple way for statistical tables of just about any probability distribution of interest.
- In R, each distribution has a name prefixed by a letter indicating whether a probability, quintile, density function or random value is required.

❖ R allows for the calculation of:

- ✓ Probabilities (including cumulative)
- ✓ The evaluation of probability density/mass functions
- ✓ quintile, and
- ✓ The generation of pseudo-random variables following a number of common distributions.

❖ Therefore, R is useful to provide a comprehensive set of statistical tables.

Cont…

- The following table gives examples of various function names in R along with additional arguments

distribution	R name	arguments
normal	`norm`	mean, sd
chi-squared	`chisq`	df, ncp
F	`f`	df1, df2, ncp
Student's t	`t`	df, ncp
exponential	`exp`	rate
log-normal	`lnorm`	meanlog, sdlog
logistic	`logis`	location, scale
Poisson	`pois`	lambda
multinomial	`multinom`	size, prob
binomial	`binom`	size, prob

Cont…

- For each distribution, R provides the following four commands:
- *dxxx:* density function of the xxx distribution
- *pxxx:* distribution function of the xxx distribution ('p' for probability) CDF
- *qxxx*: quintile function of the xxx distribution
- *rxxx:* random number generator for the xxx distribution where 'xxx' is the R name of the distribution.
- for example for normal distribution: normal

Cont…

Densities

- The density for a continuous distribution is a measure of the relative probability of "getting a value close to x". The probability of getting a value in a particular interval is the area under the corresponding part of the curve.
- **Cumulative distribution functions**
- The cumulative distribution function describes the probability of "hitting" x or less in a given distribution. The corresponding R functions begin with a 'p' (for probability) by convention

Quantiles

- The quantile function is the inverse of the cumulative distribution function.
- The p-quantile is the value with the property that there is probability p of getting a value less than or equal to it. The median is by definition the 50% quantile.

Cont…

Example1:

>dbinom(3,size=10,prob=0.25) # P(X=3) for X~Bin(n=10, p=0.25)

>dpois(0:2, lambda=4) # P(X=0), P(X=1), P(X=2) for X ~ Poisson(4)

>pbinom(3,size=10,prob=.25) # P(X < 3) in the above distribution

>pnorm(12,mean=10,sd=2) # P(X < 12) for X~N(mu = 10, sigma =2)

> qnorm(.75,mean=10,sd=2) # 3rd quartile of N(mu = 10,sigma = 2)

> qchisq(.10,df=8) # 10th percentile of $x^2(8)$

> qt(.95,df=20) # 95th percentile of t(20)

Cont…

>rnorm(100) # simulate(generate) 100 standard normal RVs

> 2*pt(-2.43, df = 13) # 2-tailed p-value for t distribution

>qf(0.01, 2, 7, lower.tail = FALSE) # upper 1% point for an F(2, 7) distribution

- Birth Day Paradox (BDP), conducted on 23 persons to have 50-50 chance that two or more of them have the same birth day from 365 days.

>*Qbirthday(prob =0.5,classes=365,coincident=2)*

> *pbirthday(23, classes = 365, coincident = 2)*

- Arguments
  - *classes:* How many distinct categories the people could fall into
  - *prob:* The desired probability of coincidence
  - *n*:The number of people
  - *coincident:* The number of people to fall in the same category

Cont…

Example2

- To find upper 5% point for an F(3,20) distribution, can be written in R

```
>qf(0.05,3,20, lower.tail=F)
```

- lower.tail=F this is the argument that the R force to do the upper tail value.
- By changing lower.tail =T we can have the lower 95% point with the similar degree of freedom.
- In the normal distribution we can also have tabular values from R.
- To find Z value of 2.15, we can write

```
1-pnorm(2.15)
```

Cont…

Example 2.3: Some biochemical measure in healthy individuals is well described by a normal distribution with mean of 132 and s.d of 13. then if a patient has a value of 160, by what probability does this can happen?

>1-pnorm(160, mean=132,sd=13)

[1] 0.01562612

Only about 1.5% of the general population that has that value or higher.

The function pnorm returns the probability of getting a value smaller than its first argument in a normal distribution with the given mean and standard deviation.

Random sampling

- In R, we can simulate different situations with the sample() function.

Example 2.4: if we want to pick five numbers at random from the set 40 numbers (1:40), then we can write

>sample(1:40,5)

- The default behavior of sample is sampling with out replacement. The sample will not contain the same number twice.

Cont…

- If we want sampling with replacement, then we need to add the argument, replace =T

>sample(c("H","T"), 10, replace =T)

- We can also generate random numbers from normal distribution by specifying the value of mean and standard deviation.

>rnorm(10, mean=7, sd=5) # this can generate 10 random numbers with mean of 7 and standard deviation 5

Cont…

Examining the distribution of a set of data

- Given a (univariate) set of data we can examine its distribution in a large number of ways. The simplest is to examine the numbers.
- Two slightly different summaries are given by
- **Summary** (is a generic function used to produce result summaries of the results of various model fitting functions, The function involves particular methods which depend on the class of the first argument.) and

  >summary(var)
- **fivenum** (the Tukey Five-Number Summaries, minimum, lower hinge, median, upper-hinge, maximum).

  > fivenum(var)

Example 2.5

```
>attach(faithful)
>summary(eruptions)
> fivenum(eruptions)
>stem(eruptions)
>hist(eruptions)       ## make the bins smaller, make a plot of density
>hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE)
 > lines(density(eruptions, bw=0.1))
> rug(eruptions)                    # show the actual data points
```

Cont…

- And also we can examine by displaying the data through the following graphs:

```
>stem(var)                  # Steam and leaf of the var.
>hist(var)                  # Default histogram of var.
>boxplot(var)               # a box plot of var
 >plot(ecdf(var))                    #the empirical cumulative
                                     #distribution function of var.
>x <- rt(250, df = 5)                # A random sample of size
                                     # 250 from t distribution with 5 df
>qqnorm(var)                # QQ plot  for normality of var.
> qqline(var)               #make a line on the above QQ plot
```

Cont…

- We can make a Q-Q plot against the generating distribution by

>qqplot(qt(ppoints(250), df = 5), x, xlab ="Q-Q plot for t dsn")

>qqline(x)

- **Formally**, R provides the **Shapiro-Wilk test** and the **Kolmogorov Smirnov test** to examine whether the given data follows a normal distribution or not.

>shapiro.test(var) *#Shapiro-Wilk test*

>ks.test(var, "pnorm", mean = mean(var), sd = sqrt(var(var)))
#Kolmogorov Smirnov test

Simulating the Sample Distribution of the Mean

- Tests on means are built on the assumption that the sample mean $\bar{X}$ is based on n independent observations from a population with mean μ and variance σ^2 . From linear combination theory, we have derived that, so long as the n observations are independent, $\bar{X}$ will have a mean of $\bar{X}$ and a variance of $\frac{\sigma^2}{n}$.

One- and two-sample t-tests

The t – tests are based on an assumption that data come from the normal distribution.

One sample tests

- In the one – sample case we thus have data x1, x2,… xn assumed to be independent realizations of random variables with distribution N(mean, variance).
- Formally we calculated

$$t = \frac{\bar{x} - \mu_0}{\delta / \sqrt{n}}$$

Two sample tests

- The two sample t – test is used to test the hypothesis that two samples may be assumed to come from distribution with the same mean.
- The theory of the two sample t – test is not very different in principle from that of the one – sample test.

Cont…

- Data are now from two groups, x11, x12, … x1n1 and x21, x22, … x2n2. which we assume are sampled from the normal distribution N(mean1, variance1) and N(mean2, variance2)
- It is desired to test the null hypothesis $\mu_1 = \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\delta^2/n + \delta^2/n}}$$

we assume s1 = s2, then the pooled S will be calculated

Cont…

- The main function that performs these sorts of tests is ***t.test().*** Its syntax is:

t.test(x, y = NULL, alternative = c("two.sided","less", "greater"), mu = 0, paired = FALSE, var.equal =FALSE, conf.level = 0.95).

Arguments:

- ***x, y:*** numeric vectors of data values. If y is not given, a one sample test is performed.
- ***alternative:*** a character string specifying the alternative hypothesis, must be one of `"two.sided"' (default), "greater" or "less". You can specify just the initial letter.
- ***mu:*** a number indicating the true value of the mean (or difference in means if you are performing a two sample test). Default is 0.
- ***paired:*** a logical indicating if you want the paired t-test (default is the independent samples test if both x and y are given).

Cont…

- ***var.equal:*** (for the independent samples test) a logical variable indicating whether to treat the two variances as being equal. If `TRUE', then the pooled variance is used to estimate the variance. If _FALSE‘ (default), then the Welch suggestion for degrees of freedom is used.
- ***conf.level:*** confidence level (default is 95%) of the interval estimate for the mean appropriate to the specified alternative hypothesis.
- Note that from the above, t.test() not only performs the hypothesis test but also calculates a confidence interval. However, if the alternative is either a “greater than” or “less than” hypothesis, a lower (in case of a greater than alternative) or upper (less than) confidence bound is given.

Cont…

Example 2.6: Test the hypotheses that the average height content of containers of certain lubricant is 10 liters if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Use the 0.01 level of significance and assume that the distribution of contents is normal.

>x=c(10.2,9.7,10.1,10.3,10.1,9.8,9.9,10.4, 10.3,9.8)

>t.test(x, mu = 10, conf.level = 0.99)

The output of the above command will be:

One Sample t-test

data: x

t = 0.7717, df = 9, p-value = 0.46

alternative hypothesis: true mean is not equal to 10

99 percent confidence interval: 9.807338, 10.312662 sample estimates:

mean of x

10.06

Cont…

Example 2.7: Consider the following sets of data on the latent heat of the fusion of ice (cal/gm) from Rice.

Method A: 79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02 80.00 80.02

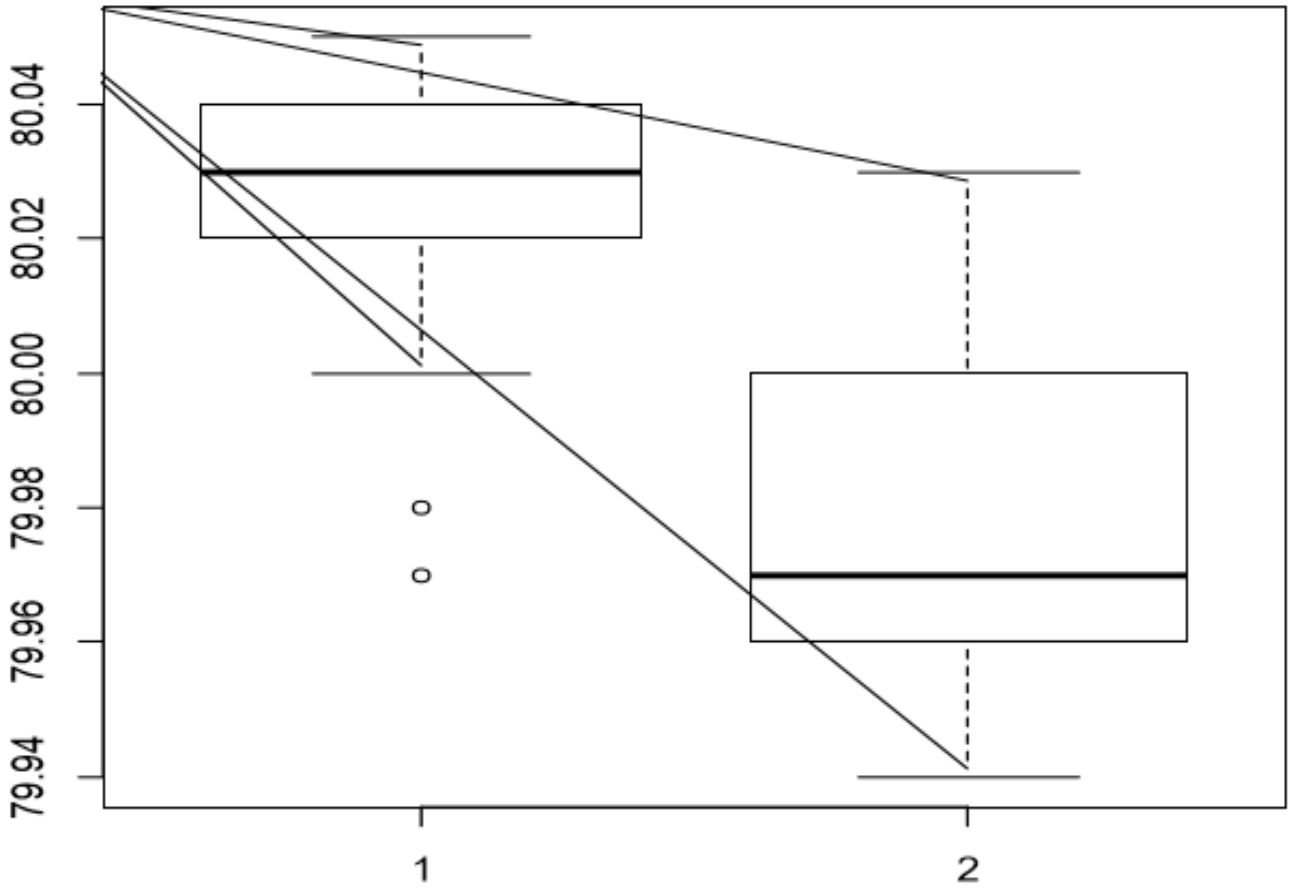
Method B: 80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97

- Box plots provide a simple graphical comparison of the two samples.

```
>A=c(79.98,80.04,80.02,80.04,80.03,80.03,80.04,79.97,80.05,80.03,80.02,80.00,80.02)
>B=c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97)
>boxplot(A,B)
```

cont…

which indicates that the first group tends to give higher results than the second.



Cont…

- To test for the equality of the means of the two samples, we can use an unpaired t-test by:

> t.test(A, B)

- This will give you the following output:

Welch Two Sample t-test

- *data: A and B t = 3.2499, df = 12.027, p-value = 0.006939*

alternative hypothesis: true difference in means is not equal to 0(zero)

95 percent confidence interval: 0.01385526, 0.07018320

sample estimates:

mean of x mean of y

80.02077 79.97875

Which indicate a significant difference, assuming normality. By default the R function does not assume equality of variances in the two samples.

Cont…

We can use the F test to test for equality in the variances, provided that the two samples are from normal populations.

>var.test(A, B) F test to compare two variances data:

A and B

F = 0.5837, num df = 12, denom df = 7, p-value = 0.3938

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.1251097 2.1052687

sample estimates:

ratio of variances

0.5837405

which shows no evidence of a significant difference, and so we can use the classical t-test that assumes equality of the variances.

Cont…

>t.test(A, B, var.equal=TRUE)

Two Sample t-test data: A and B

t = 3.4722, df = 19, p-value = 0.002551

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.01669058 0.06734788

sample estimates:

mean of x mean of y

80.02077 79.97875

All these tests assume normality of the two samples. The two-sample Wilcoxon (or MannWhitney) test only assumes a common continuous distribution under the null hypothesis.

> wilcox.test(A, B)

Wilcoxon rank sum test with continuity correction

data: A and B W = 89, p-value = 0.007497

alternative hypothesis: true location shift is not equal to 0

Cont…

The paired t test

Paired tests are used when there are two measurements on the same experimental unit. Their theory is essentially based on taking differences and thus reducing the problem to that of a one-sample test

> t.test(pre, post, paired=T)

Exercise2.1: the recovery time (in days) is measured for 10 patients taking a new drug and for 10 different patients taking a placebo. We wish to test the hypothesis that the mean recovery time for patients taking the drug is less than fort those taking a placebo (under an assumption of normality and equal population variances). The data are:

With drug: 15, 10, 13, 7, 9, 8, 21, 9, 14, 8

Placebo: 15, 14, 12, 8, 14, 7, 16, 10, 15, 12

cont...

Answer

```
 > drug <- c(15, 10, 13, 7, 9, 8, 21, 9, 14, 8)
> plac <- c(15, 14, 12, 8, 14, 7, 16, 10, 15, 12)
> t.test(drug, plac, alternative ="less", var.equal = T)
Two Sample t-test
data: drug and plac
t = -0.5331, df = 18, p-value = 0.3002
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf 2.027436
sample estimates:
mean of x mean of y
11.4 12.3
```

Cont…

Exercise 2.2: an experiment was performed to determine if a new gasoline additive can increase the gas mileage of cars. In the experiment, six cars are selected and driven with and without the additive. The gas mileages (in miles per gallon, mpg) are given below.

Car	1	2	3	4	5	6
mpg w/ additive:	24.6	18.9	27.3	25.2	22.0	30.9
mpg w/o additive:	23.8	17.7	26.6	25.1	21.6	29.6

Solution:

Since this is a paired design, we can test the claim using the paired t–test (under an assumption of normality for mpg measurements). This is performed by:

```
add <-c(24.6, 18.9, 27.3, 25.2, 22.0, 30.9)
>noadd <-c(23.8, 17.7, 26.6, 25.1, 21.6, 29.6)
```

Cont…

>t.test(add, noadd, paired=T, alt = "greater")

The output is given as follows: Paired t-test

data: add and noadd t = 3.9994, df = 5, p-value = 0.005165 **alternative hypothesis:** true difference in means is greater than 0

95 percent confidence interval:

0.3721225 Inf

sample estimates:

mean of the differences

0.75