Psychol. Statistics using R

TAIPEI MEDICAL UNIVERSITY

臺北醫學大學

TAIPEI MEDICAL UNIVERSITY

Categorical Testing

Changwei W. Wu, Ph.D.

Graduate Institute of Mind, Brain and Consciousness
Research Center of Brain and Consciousness
Taipei Medical University

Statistics

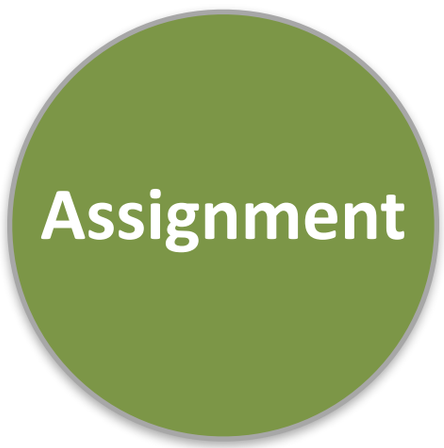
1. Deal with Proportions

- Binomial test for percentage or rates
- Normal approximation

2. Deal with Counts

- Chi-square test
- Fisher's exact test

3. [R] Hands-on Practices



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

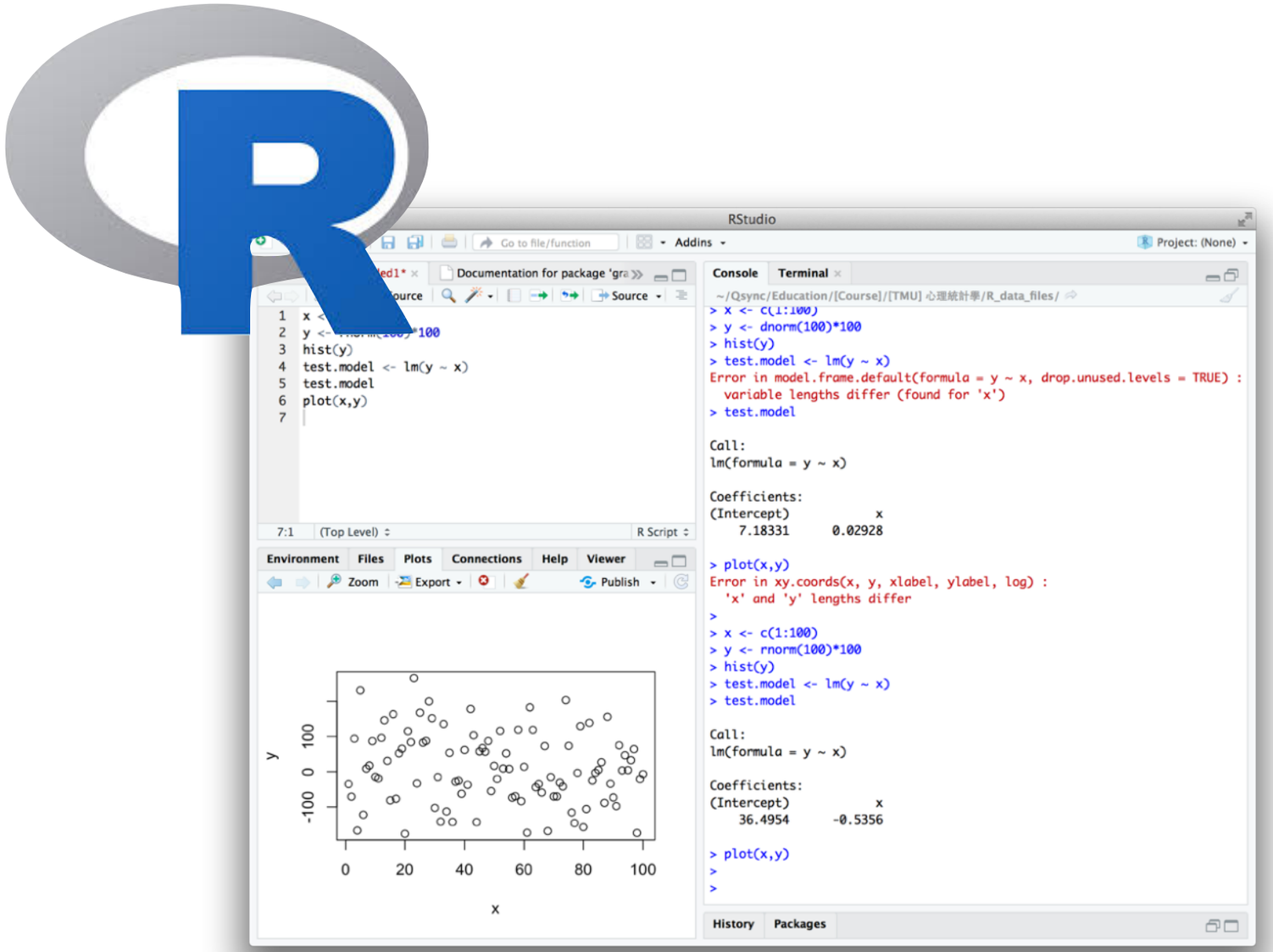
Types of Variables

- **Numeric variables**
  - Equal-interval variables (interval scales)
    - Continuous variable: e.g., GPA; height
    - **Discrete counts**: e.g., the number of time visiting dentist
    - **Proportions**: e.g., percentage, rate
  - Rank-order variables (ordinal/discrete scales)
    - e.g., order of finishing a race; birth order of children
    - Physical activity level (low, moderate and high)
- **Nominal variables**
  - Gender (male, female)
  - Ethnicity (Caucasian, African American, Asian and Hispanic)
  - Profession (surgeon, doctor, nurse, dentist)

Binomial distributions

Normal approximation

① BINOMIAL DISTRIBUTIONS (PROPORTIONS)



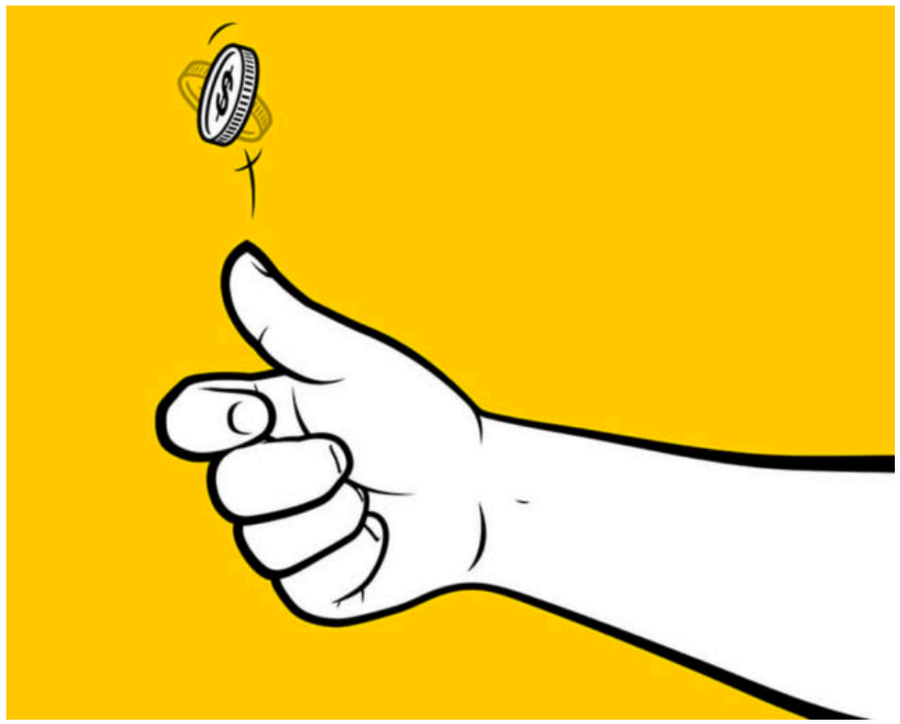
TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Binomial Distribution

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

Success Fail

50% 50%



probability with many trials

- ***dbinom(x, n, p)***

THEOREM 3.2. Let X be a binomial random variable with parameters n and p. Then $\mu = E(X) = np$ and $\sigma_X^2 = \mathrm{Var}(X) = np(1-p)$.

Binomial Distribution

DEFINITION 3.6. The probability density function of a **binomial random variable** with n trials and probability p of success on any trial is

Success Failure

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

The binomial random variable probability density function is characterized by the two parameters n, the number of trials or sample size, and p, the probability of success on a trial.

1. A fixed number n of trials are carried out.

2. The outcome of each trial can be classified in precisely one of two mutually exclusive ways termed "success" and "failure." The term "binomial" literally means two names.

3. The probability of a success, denoted by p, remains constant from trial to trial. The probability of a failure is $1 - p$.

4. The trials are independent; that is, the outcome of any particular trial is not affected by the outcome of any other trial.

THEOREM 3.2. Let X be a binomial random variable with parameters n and p. Then $\mu = E(X) = np$ and $\sigma_X^2 = \mathrm{Var}(X) = np(1-p)$.

Binomial Distribution

EXAMPLE 3.9. A particular strain of inbred mice has a form of muscular dystrophy that has a clear genetic basis. In this strain the probability of appearance of muscular dystrophy in any one mouse born of specified parents is $\frac{1}{4}$. If 20 offspring are raised from these parents, find the following probabilities.

(a) Fewer than 5 will have muscular dystrophy;

(b) Five will have muscular dystrophy;

(a) To determine the probability that fewer than 5 will have muscular dystrophy, use $n = 20$ and $p = 0.25$. Then $P(X < 5) = F(4)$. In Table C.1 we look up the column corresponding to $n = 20$ and $p = 0.25$. The row denoted by the 4 in the margin gives the CDF for 4 successes in 20 trials, i.e., the probability of 4 or fewer successes in 20 trials:

$$P(X < 5) = F(4) = 0.4148.$$

➡ **Use R code to calculate**

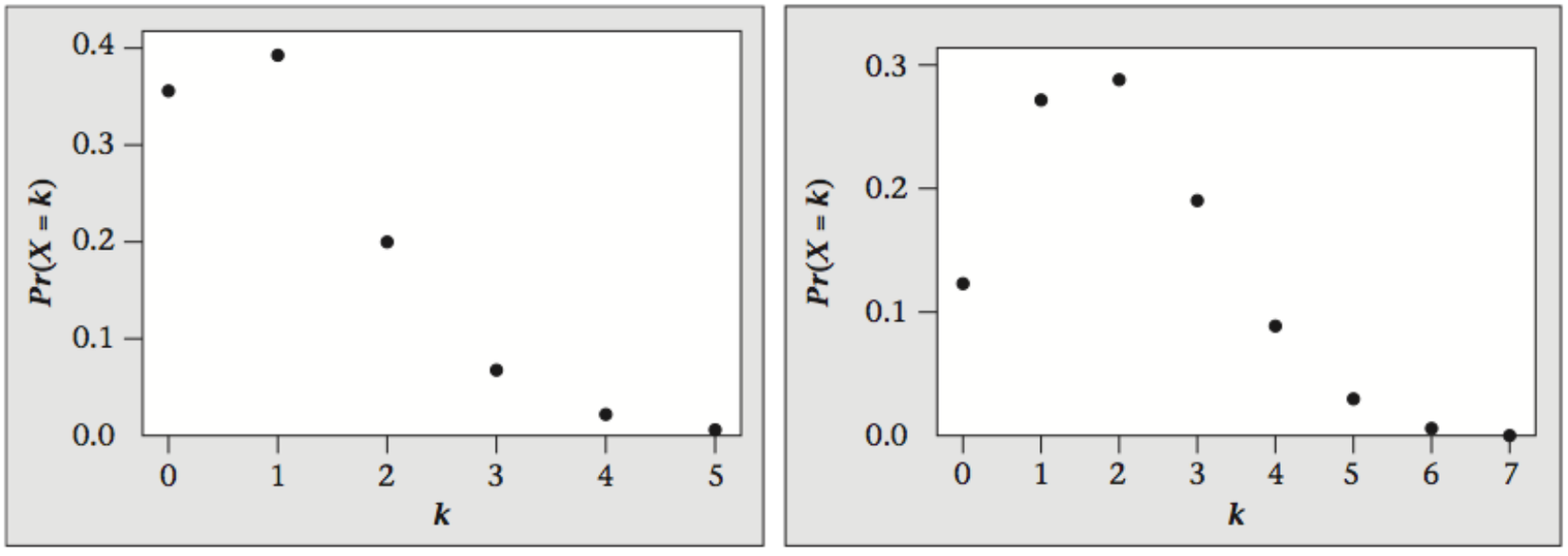
(b) To determine the probability that 5 will have muscular dystrophy, use

$$P(X = 5) = f(5) = F(5) - F(4) = 0.6172 - 0.4148 = 0.2024.$$

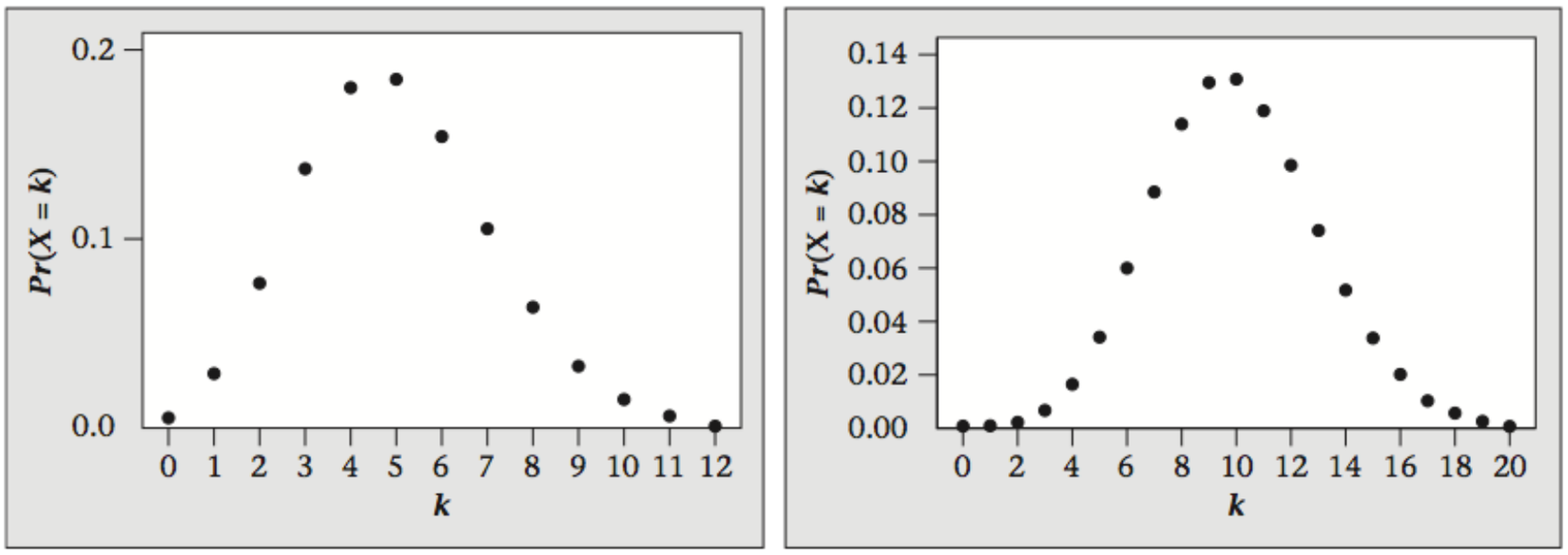
Here again we could evaluate the density function directly, but a simple subtraction of adjacent CDF values is easier and quicker than calculating

$$f(5) = \frac{20!}{5!(20-5)!}\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^{20-5} = 0.2024.$$

Normal Approximation (Z-dist.)



(a) $n = 10, p = .1$ (b) $n = 20, p = .1$



(c) $n = 50, p = .1$ (d) $n = 100, p = .1$

Normal Approximation (Z-dist.) & Continuity Correction

FORMULA 3.1. Let X be a binomial random variable with parameters n and p. For large values of n, X is approximately normal with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. The approximation is acceptable for values n and p such that $np > 5$ and $n(1-p) > 5$. In this situation,

$$F_B(X) \approx F_N\left(\frac{X + 0.5 - np}{\sqrt{np(1-p)}}\right),$$

where F_B and F_N are the cumulative binomial and normal distributions, respectively.

Continuity correction

➡ Difference btw *Discrete* & *Continuous*

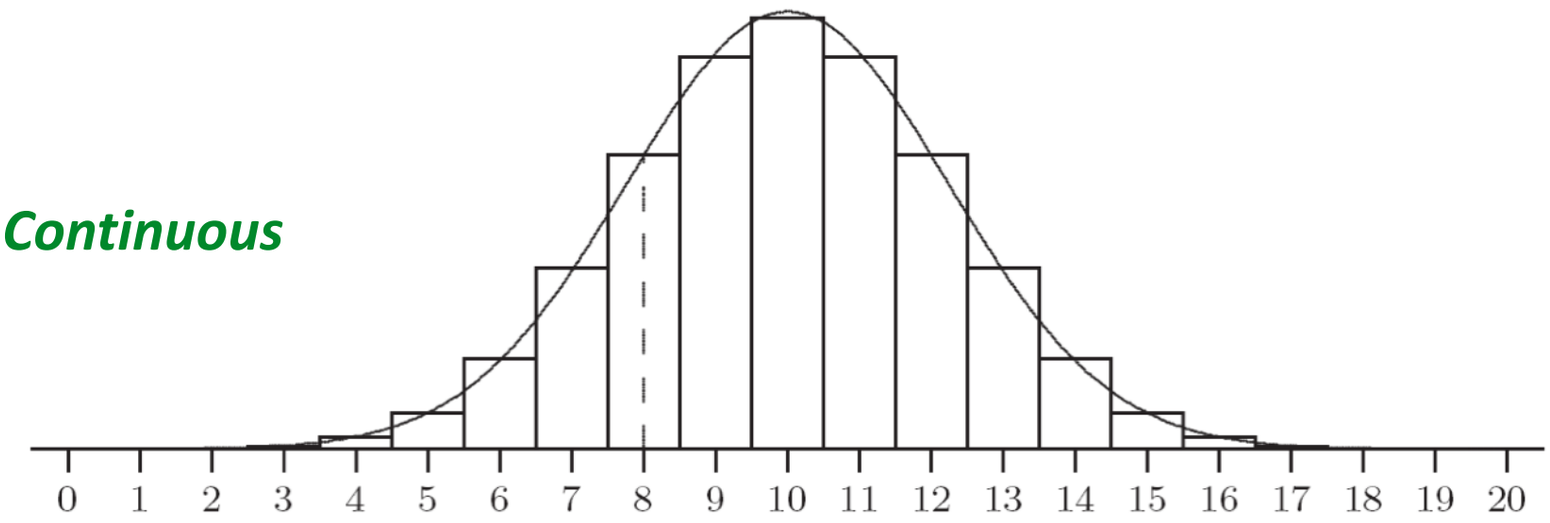


FIGURE 3.17. The binomial distribution with $n = 20$ and $p = 0.5$ is closely approximated by a normal distribution with $\mu = np = 10$ and $\sigma = \sqrt{np(1-p)} = \sqrt{5}$.

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

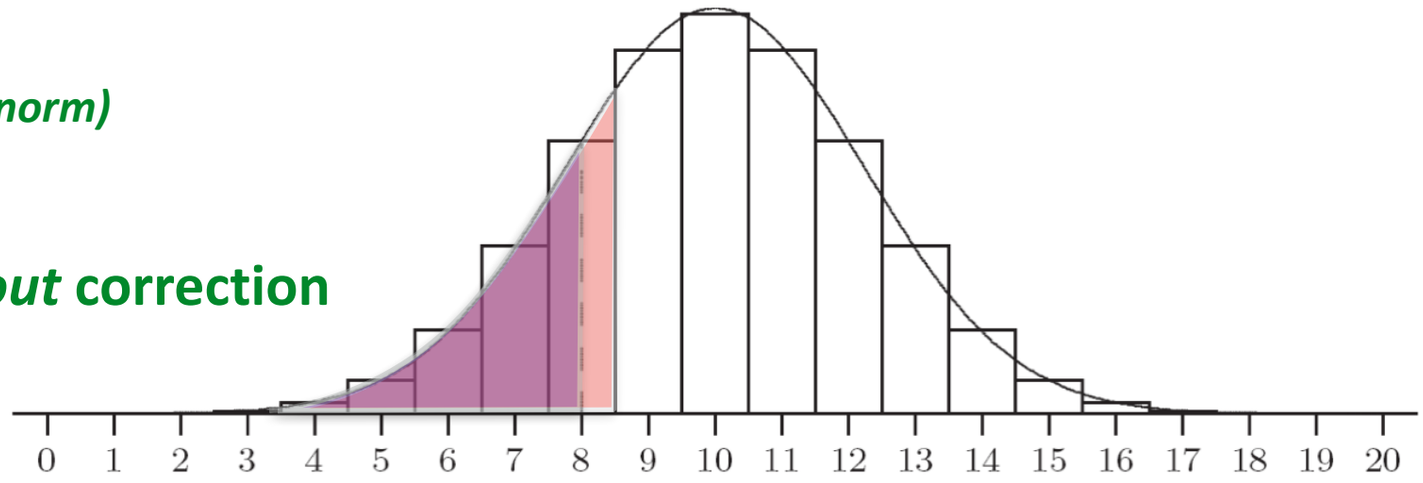
Normal Approximation (Z-dist.) & Continuity Correction

EXAMPLE 3.18. Return to Example 3.9 with a genetic form of muscular dystrophy in mice. This example was solved using binomial probabilities with $p = 0.25$ and $n = 20$. Suppose a larger sample of progeny are generated and we would like to know the probability of fewer than 15 with muscular dystrophy in a sample of 60.

Use normal approximation *(pnorm)*

- ➡ **Binomial distri.**
- ➡ **Normal Approx. *With* & *Without* correction**

Practice



On the other hand, the normal approximation can be done much more quickly as

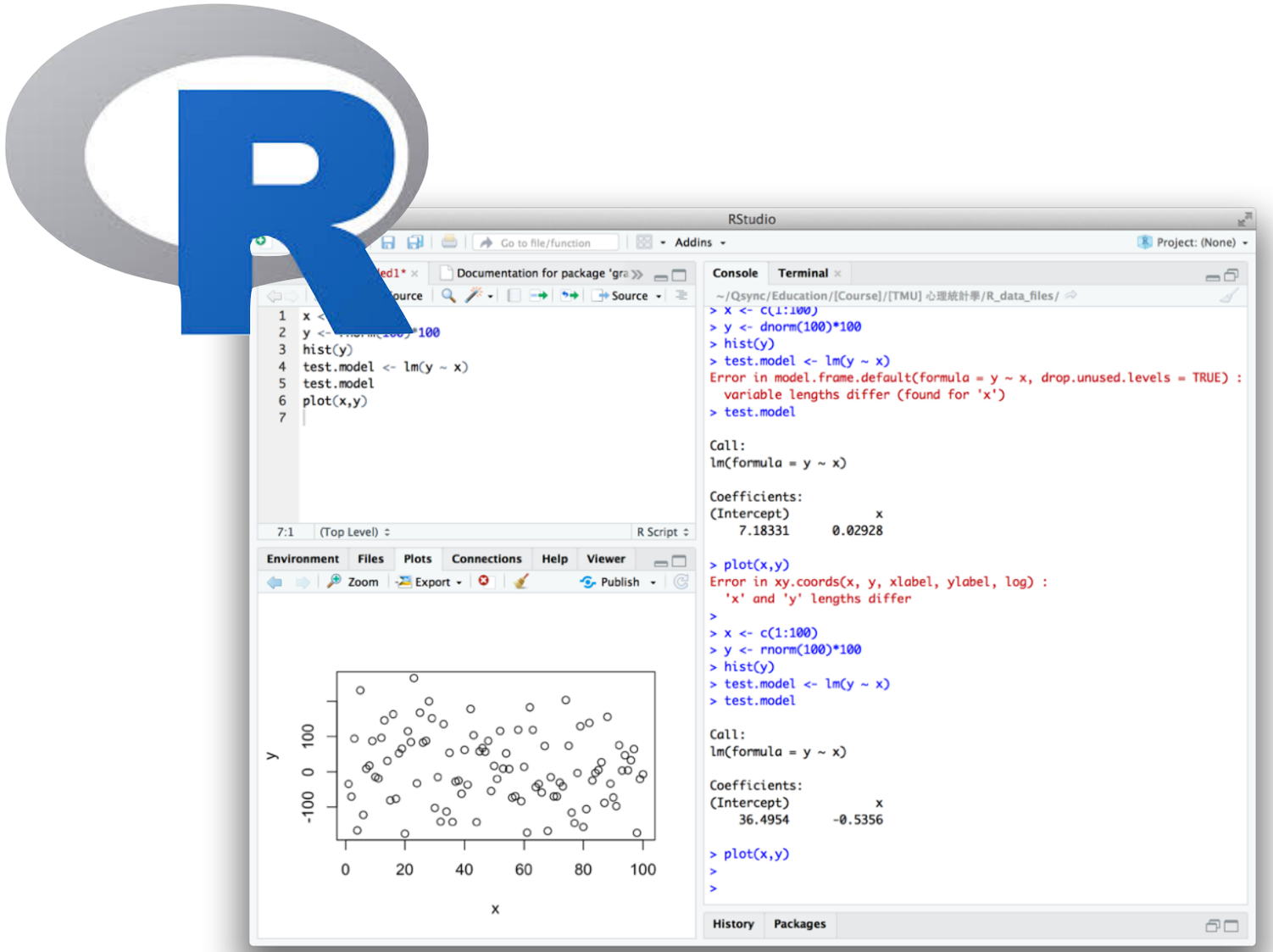
$$P(X < 15) = F_B(14) \approx F_N\left(\frac{14.5 - \mu}{\sigma}\right) = F_N\left(\frac{14.5 - 15}{3.35}\right)$$

$$= F_N(-0.15) = 0.4404$$

using Table C.3. So with relatively little effort, the actual CDF for a binomial with $n = 60$ and $p = 0.25$ is reasonably well-approximated by the normal distribution using $\mu = 15$ and $\sigma = 3.35$. But why was 14.5 instead of 14 used in this approximation? This 0.5 correction term was used because we were estimating a discrete distribution using a continuous one. For this reason, it is called a **continuity correction**.

Foundation & Installation

② ONE-SAMPLE TEST FOR PROPORTIONS



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Binomial for Proportion (One-sample)

- **Assumptions**
  - An independent random sample of size **n** is drawn from the population.
  - The population is composed of 2 mutually exclusive categories.
  - p is the actual proportion of the population in the first of the 2 categories.
  - p_0 is the hypothesized value of p. (pre-defined)

- **Hypotheses**
  - [Two-tailed] H_0: $p = p_0$ and H_A: $p \neq p_0$
  - [Left-tailed] H_0: $p \geq p_0$ and H_A: $p < p_0$
  - [Right-tailed] H_0: $p \leq p_0$ and H_A: $p > p_0$

- **Test statistics**
  - X is a binomial random variable (the number of "success") with parameters n and $p = p_0$.

DEMO

Growth Ring

EXAMPLE 11.2. The severe drought of 1987 in the U.S. affected the growth rate of established trees. It is thought that the majority of the trees in the affected areas each have a 1987 growth ring that is less than half the size of the tree's other growth rings. A sample of 20 trees is collected and 15 have this characteristic. Do these data support the claim?

15 "success" out of 20 samples

Measure: *p* = proportion of trees have this property

- **Set up Hypothesis**:

① Hypothesis

H_0: $p_{(\text{trees in this area})} \leqq 0.5$

H_a: $p_{(\text{trees in this area})} > 0.5$ ($\because$majority)

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

DEMO

Growth Ring

② Assumption

independent samples?

③ Testing

General form:

- ***binom.test(x, n, p=0.5, alternative = c("two.sided", "less", "greater"))***

- **Real usage:**
  - ***binom.test(15, 20, p=0.5, alternative = "greater")***

```
        Exact binomial test

data:  15 and 20
number of successes = 15, number of trials = 20, p-value =
0.02069
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5444176 1.0000000
sample estimates:
probability of success
                  0.75
```

➡ **p = 0.02069 < (α = 0.05), Reject H_0**

DEMO

Growth Ring

④ Effect Size

```
        Exact binomial test

data:  15 and 20
number of successes = 15, number of trials = 20, p-value =
0.02069
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5444176 1.0000000
sample estimates:
probability of success
                  0.75
```

➡ p = 0.02069 < (α = 0.05), Reject H_0

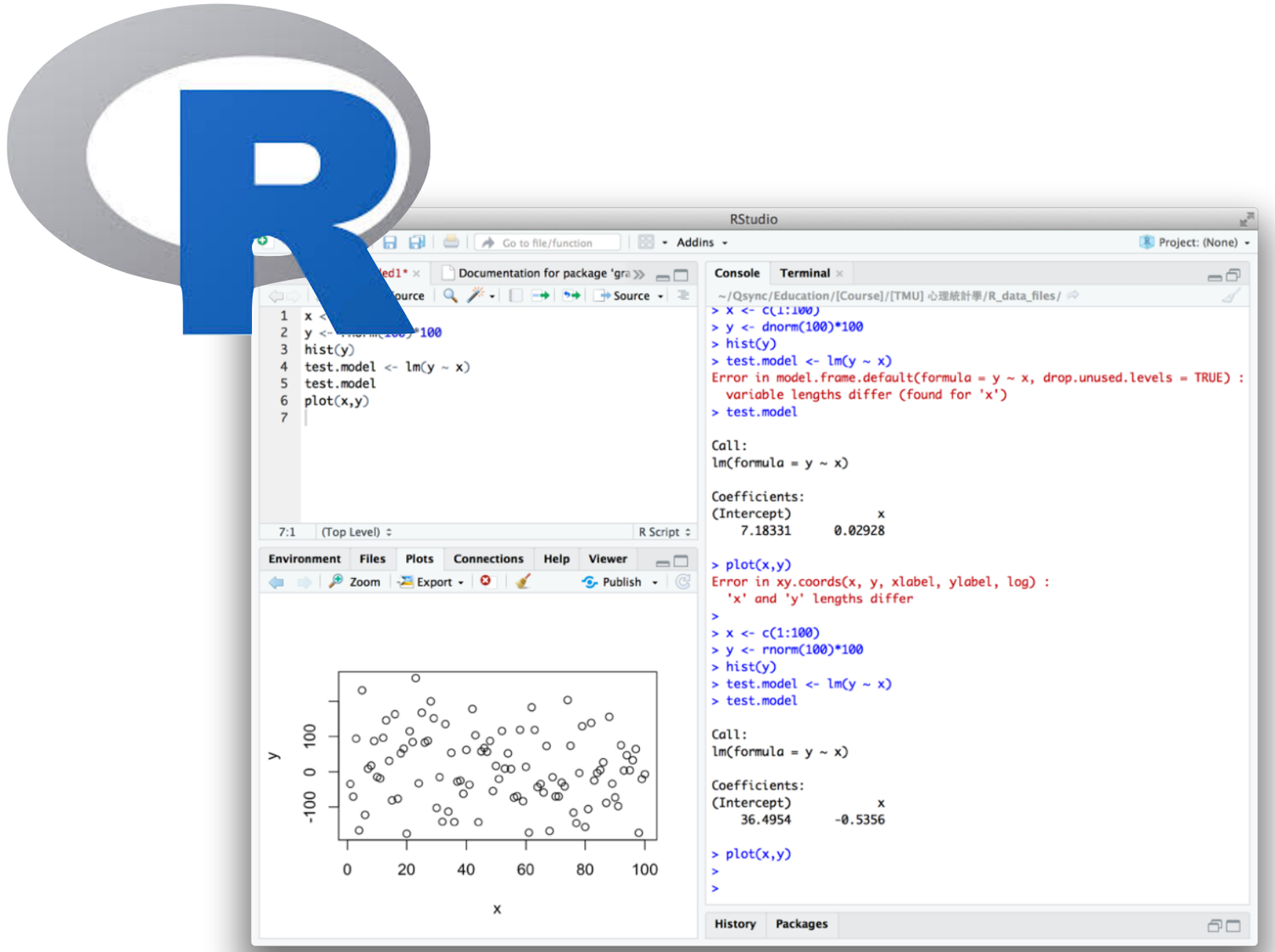
Odds Ratio (OR) = 0.75/0.5 = 1.5

⑤ Decision

- **Reporting decision:**
  There is significant evidence that the majority of trees (75%) have growth rings of 1987 less than half their usual size ($p < 0.021$).

Independent Observations

② TWO-SAMPLE TEST FOR PROPORTIONS



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Binomial for Proportion (Two-sample)

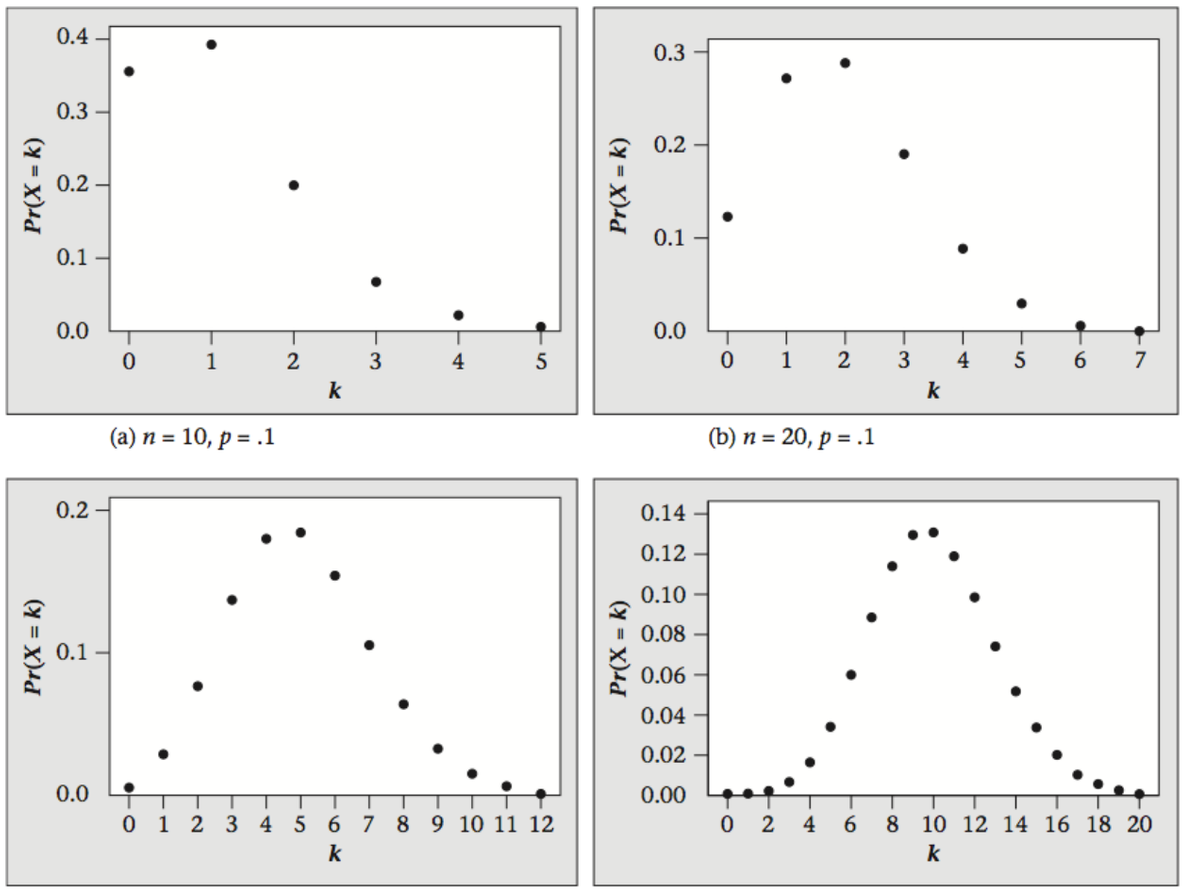
- **Normal approximation**

  - The sample size n is sufficiently large so that $np_0 > 5$ and $n(1-p_0) > 5$.

Let X continue to denote the number of successes or observations falling into the first category. Since $np_0 > 5$ and $n(1-p_0) > 5$, then, as we saw at the end of Section 3.6, X is approximately normal with mean $\mu = np_0$, variance $\sigma^2 = np_0(1-p_0)$, and standard deviation $\sigma = \sqrt{np_0(1-p_0)}$.

$$z = \frac{X-\mu}{\sigma} = \frac{X-np_0}{\sqrt{np_0(1-p_0)}}$$

$$z = \frac{\frac{X}{n}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$



(a) $n = 10$, $p = .1$ (b) $n = 20$, $p = .1$

(c) $n = 50$, $p = .1$ (d) $n = 100$, $p = .1$

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

Binomial for Proportion (Two-sample)

- **Normal approximation**
  - pooled variance p_c = (n1*p1 + n2*p2) / (n1 + n2)

$$z = \frac{\hat{D}}{\mathrm{SE}_{\hat{D}_c}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

- **Hypotheses**
  - [Two-tailed] H_0: $p_1 = p_2$ and H_A: $p_1 \neq p_2$
  - [Left-tailed] H_0: $p_1 \geq p_2$ and H_A: $p_1 < p_2$
  - [Right-tailed] H_0: $p_1 \leq p_2$ and H_A: $p_1 > p_2$

DEMO

Death rate after Stenting

EXAMPLE 11.6. A stent is a small wire mesh tube, often inserted in an artery, that acts as a scaffold to provide support to keep the artery open. Stents are a common therapy used to keep open totally occluded arteries in coronary patients. They may be implanted even several days after a heart attack under the assumption that any procedure that increases blood flow will lead to an improvement in survival rate.

A study was carried out to determine whether the time when stents are implanted in patients' arteries changes the effectiveness of the treatment. In the study 2166 stable patients who had had heart attacks in the previous 3 to 28 days and had a total occlusion of an artery related to this attack were randomly assigned to two groups. Group 1 consisted of 1082 patients who received stents and optimal medical therapy, while Group 2 consisted of 1084 patients who received medical therapy alone with no stents. The results were that 171 of the patients in Group 1 and 179 patients in Group 2 had died by the end of four years.

Researchers had expected to find a reduction in death rate for patients receiving stents under the conditions described. Was there any evidence for this?

G1: 171 out of 1082 | G2: 179 out of 1084

Measure: ***sampling proportions is not the same***

An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

DEMO

Death rate after Stenting

- **Hypothesis**:

H_0: $p_{(death\_G1)} \geqq p_{(death\_G2)}$

H_a: $p_{(death\_G1)} < p_{(death\_G2)}$

death <- c(171,179)
total <- c(1082,1084)

③ Testing

General form:

- *prop.test(success, trials, alternative = “two.sided”, correct = T)*

- **Usage:**
  - *prop.test(death, total, alternative= "less", correct=F)*

```
data:  passaway out of stenting
X-squared = 0.20083, df = 1, p-value = 0.327
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01892715
sample estimates:
   prop 1    prop 2
0.1580407 0.1651292
```

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

DEMO

Death rate after Stenting

④ Effect Size

```
data:  passaway out of stenting
X-squared = 0.20083, df = 1, p-value = 0.327
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01892715
sample estimates:
   prop 1    prop 2
0.1580407 0.1651292
```

➡ $p = 0.327 > (\alpha = 0.05)$, Accept H_0

$\phi = sqrt(\chi^2/n)$

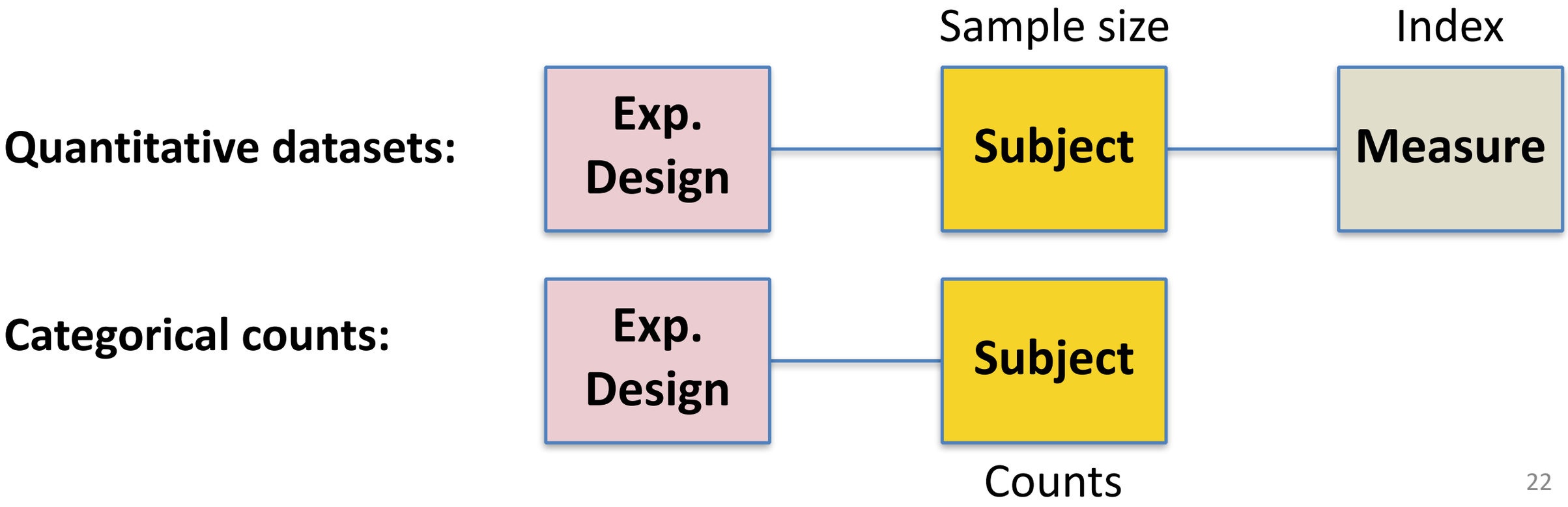
$= sqrt(0.201 / 2166) = 0.01$

⑤ Decision

- **Reporting decision:**
  There is no evidence for a significant reduction in death rate with the implantation of stents in patients after heart attack (p=0.327).

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

2. Dealing w/ ***Counts***



Chi-Squared Statistics χ^2

- A great deal of statistical information comes in the form of counts: the number of male participants, number of branches on a tree, number of days of frost…etc.

- **Contingency table**: all the events of *nominal* (categorical) variables could possibly happen.
  - ➢ Hair color, political party, sex/gender, choice of college major, religion, races, attachment style, etc.

- **Basic idea**: compare how well an *observed frequency* of people over various categories fits *expected frequency*.

Chi-Squared Statistics χ^2

Brain Imaging and Behavior
https://doi.org/10.1007/s11682-020-00364-w

ORIGINAL RESEARCH

Spontaneous thought-related network connectivity predicts sertraline effect on major depressive disorder

χ^2 *Tests*

for nominal variables

	Normal control ($n = 35$)	MDD Patient ($n = 22$)	p value
Mean age (SD)	40.1 (10.6)	40.7 (11.6)	0.6[a]
Right-handedness (%)	33 (94.2)	21 (95.5)	0.8[b]
Male (%)	12 (34.3)	5 (22.7)	0.4[b]
Mean education (SD)	15.3 (1.8)	13.0 (3.8)	0.4[c]
Mean HAM-D score (SD)	0 (0.0)	Pre: 24.5 (2.9) Post: 7.9 (5.3)	<0.001[d]
Anxiety (%)	0 (0.0)	6 (27.3)	NA

[a] Mann-Whitney U test

[b] Chi-square test

[c] Median test

[d] Paired t-test between pre- and post-treatment MDD patients

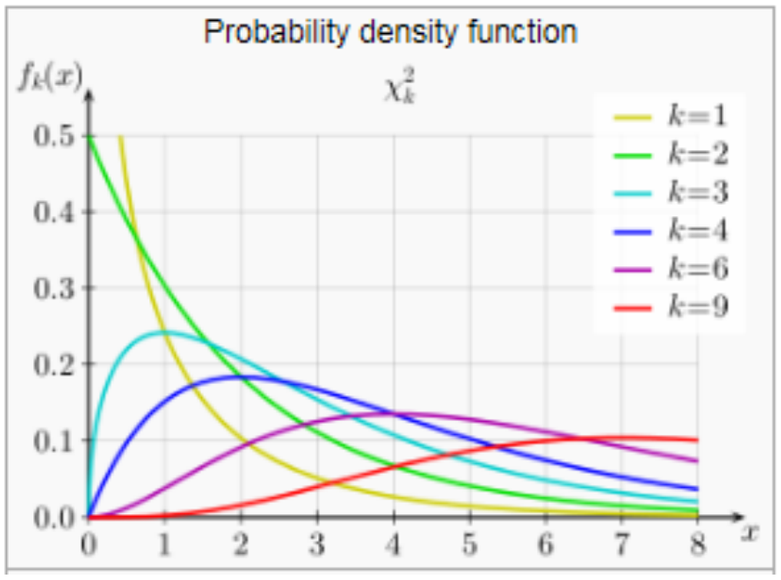
Chi-square Test

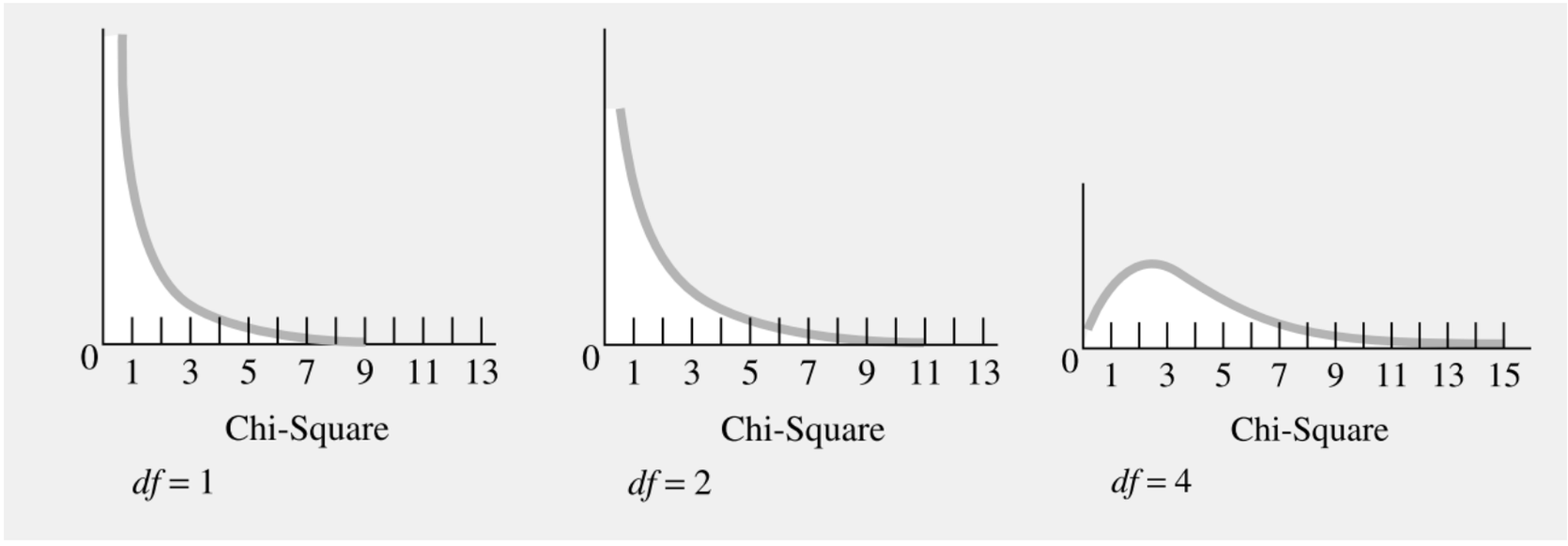
Expected value

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (X_i - \mu)^2 / \sigma^2$$

- To test the relationship between two ***categorical*** variables.
  - Check to make sure that no expected frequencies are less than 5.
  - Report the contingency table.
  - Report the χ^2 statistic, *degrees of freedom* & *odds ratio* for effect size.

chi-squared





臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Confidence Interval for Variance

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} (X_i - \mu)^2 / \sigma^2 \sim \chi_n^2 = \text{chi-square distribution with } n\,df$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 / \sigma^2 \sim \chi_{n-1}^2$$

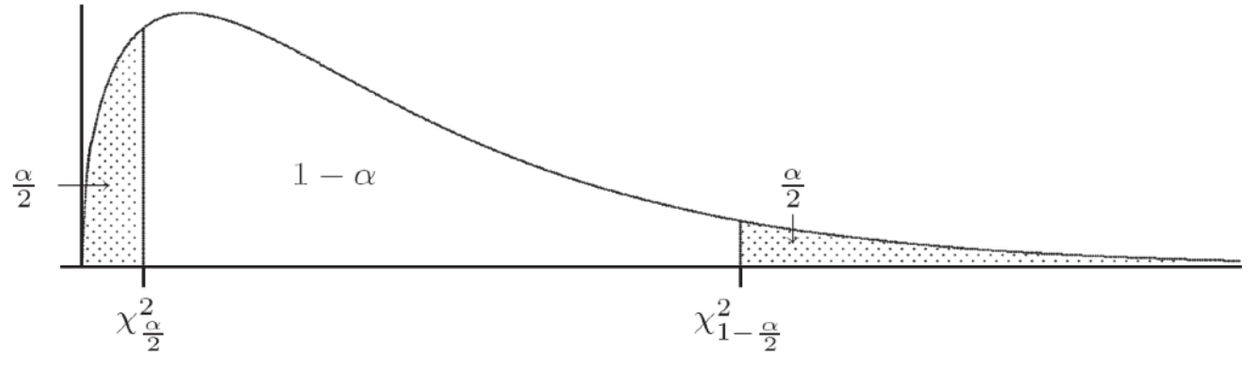
When using sample data, take off 1 DOF.

FORMULA 4.2. A $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is given by

$$C\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\right) = 1-\alpha,$$

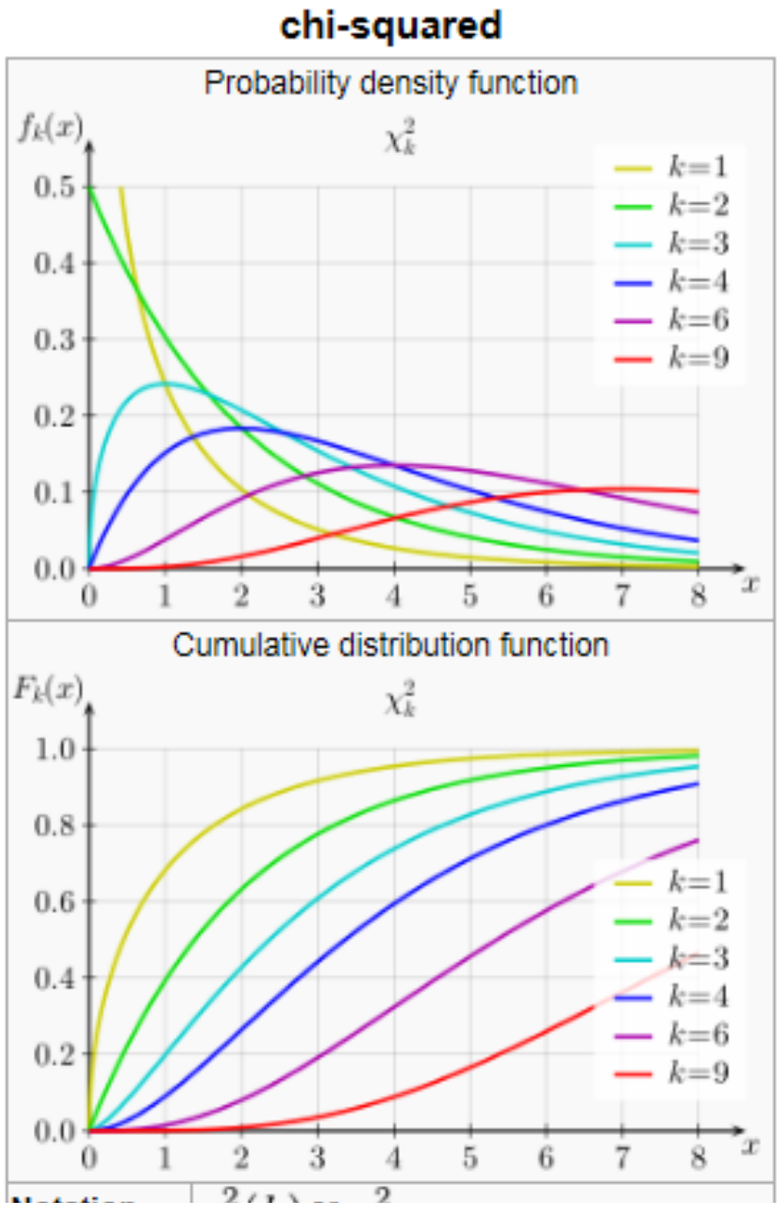
where the chi-square distribution has $n-1$ degrees of freedom. Explicitly, the interval endpoints are

$$L_1 = \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \text{and} \quad L_2 = \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}.$$



An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

Chi-Square Distribution



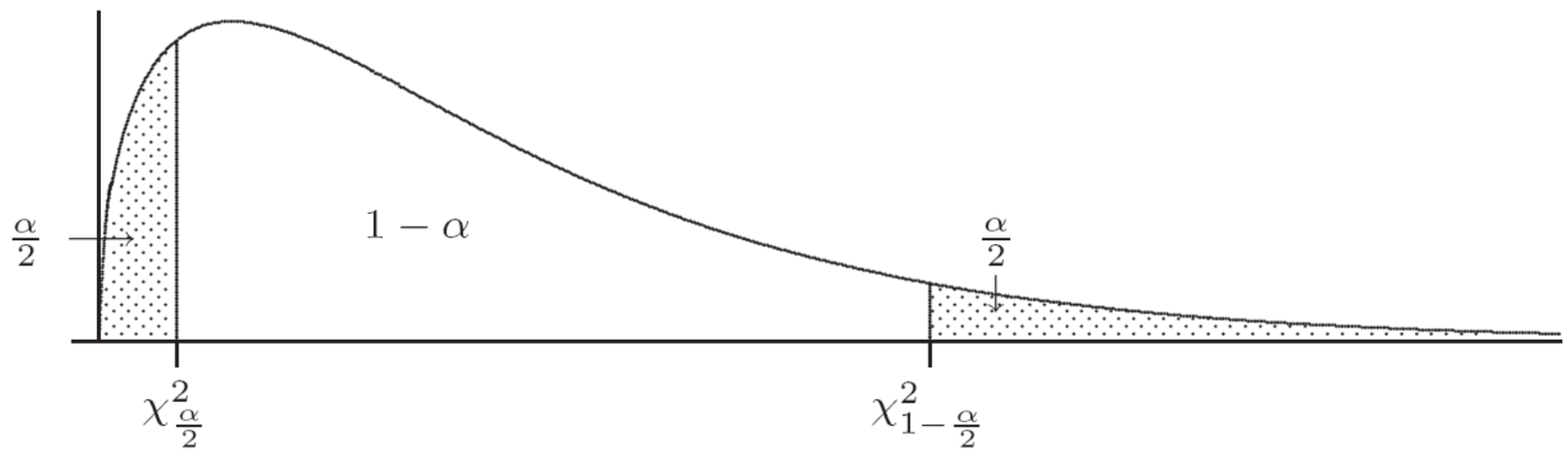
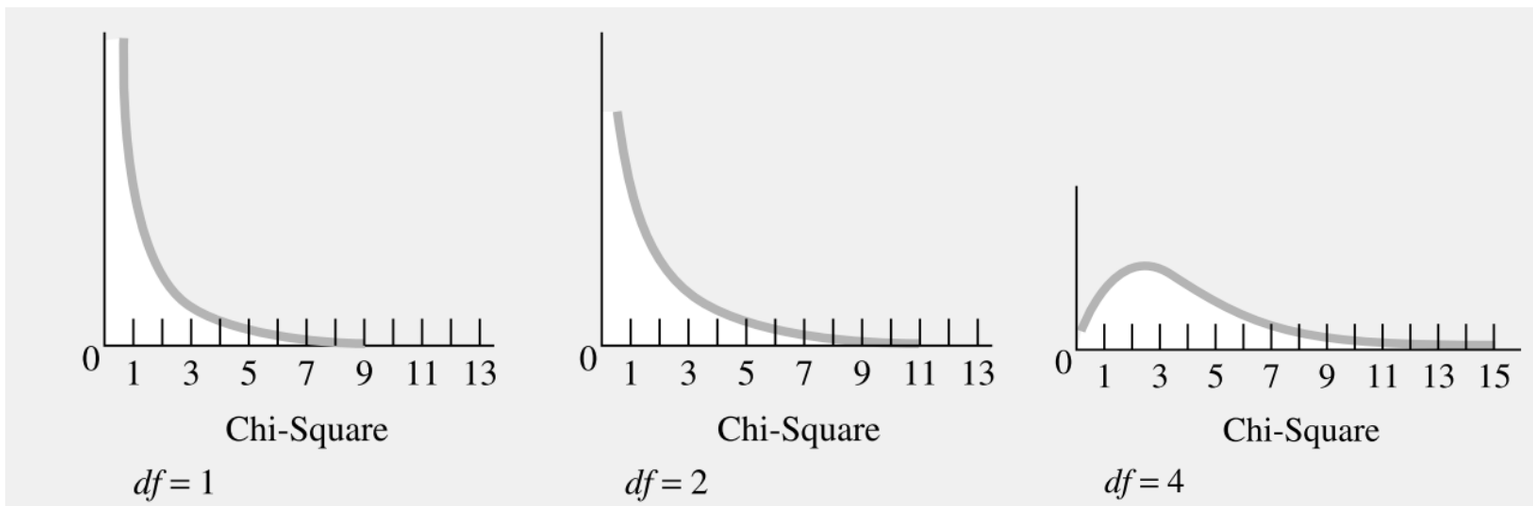


FIGURE 4.7. Locating the middle area of size $1-\alpha$ in a χ^2 distribution.

Notice: the chi-square distribution is NOT symmetric
(unlike the normal distribution).



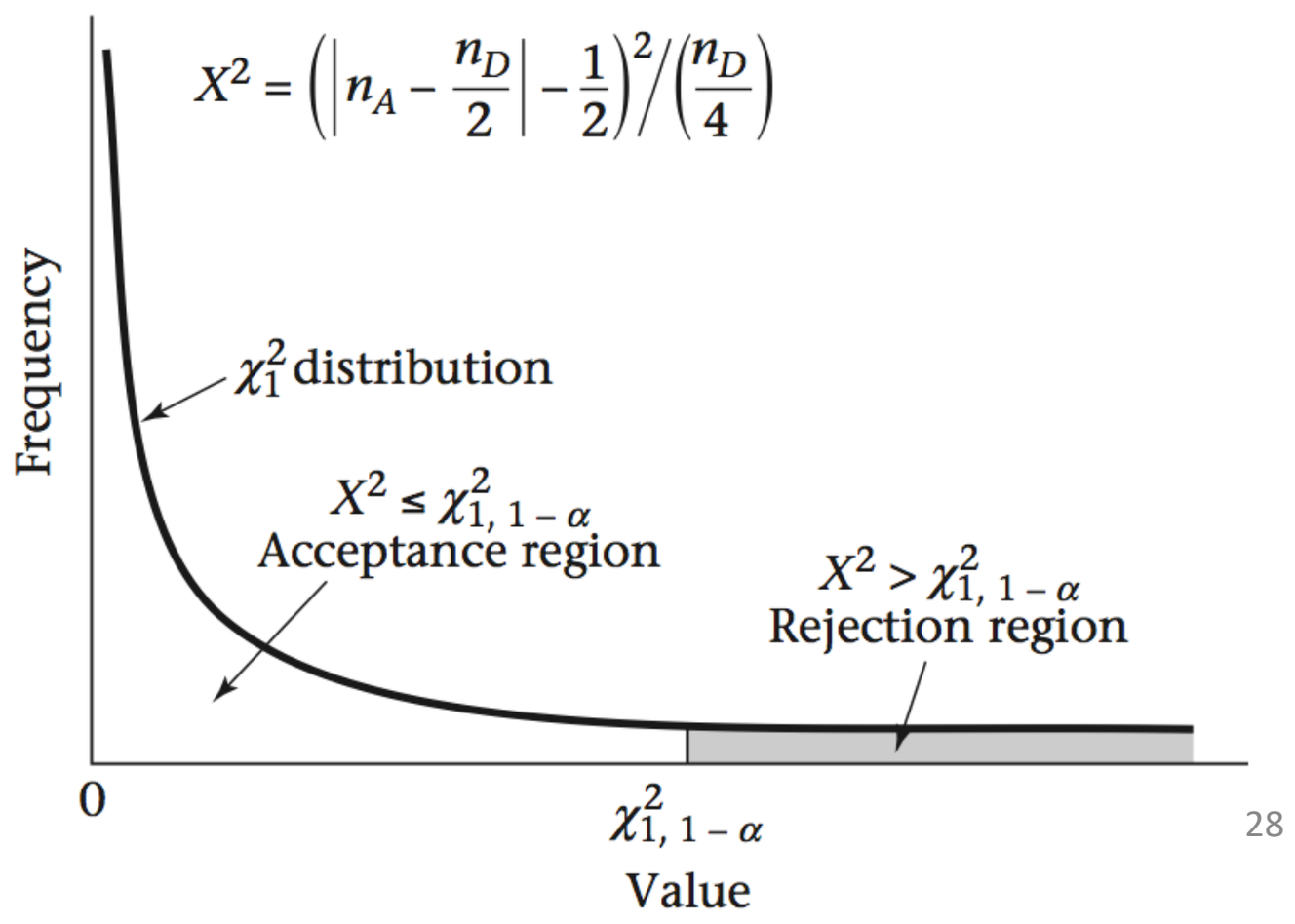
Chi-Squared Statistics χ^2

- Two types of chi-square tests:
  - ➢ **Chi-square test for goodness of fit (one-variable)**

    A chi-square test involving levels of a single categorical variable.
  - ➢ **Chi-square test for independence (two-variable)**

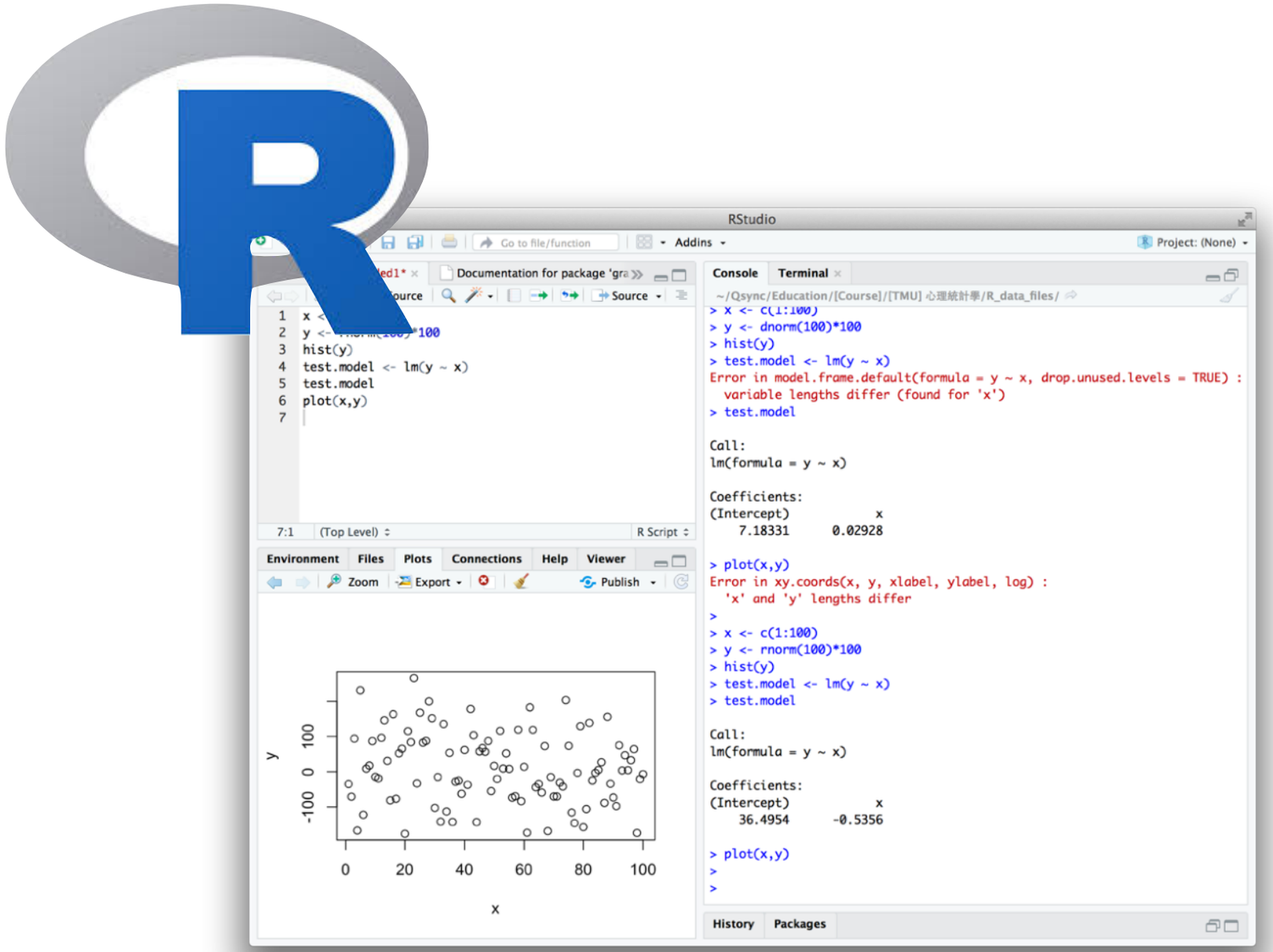
    A chi-square test when there are two categorical variables, each with several categories.

    – The test is always ***right-tailed.***



Foundation & Installation

③ TEST ON CONTINGENCY TABLES



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Chi-Square Test (Goodness of Fit)

- **Assumptions**
  - An independent random sample of size **n** is drawn from the population.
    - not for repeated-measure design (use McNemar test for paired case).
  - The population can be divided into a set of k mutually exclusive categories.
  - The expected **counts/frequencies** for each category must be specified.
  - Sample size must be sufficient for each expected value $\mathbf{E_i}$ is larger than ***5***.

- **Hypotheses**
  - H_0: The observed frequency distribution is the same as the hypothesized frequency distribution
  - H_A: The observed and hypothesized frequency distributions are different.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Chi-Square Test (Goodness of Fit)

- **Test statistics**

  –

$$\chi^2 = \sum \frac{\left(\text{observed}_{ij} - \text{model}_{ij}\right)^2}{\text{model}_{ij}} \qquad df = k - 1$$

- **Yate's continuity correction**
  - Using contingency table, Pearson's chi-square tends to increase **Type I error**.
  - Adding **-0.5** is to lower the test statistics (less significant).
  - Use the correction when the sample size is small.

$$\chi^2 = \sum \frac{(|\text{observed}_{ij} - \text{model}_{ij}| - 0.5)^2}{\text{model}_{ij}}$$

Hands-On

Inherited Flower Colors

EXAMPLE 11.8. The progeny of self-fertilized four-o'clocks were expected to flower red, pink, and white in the ratio 1:2:1. There were 240 progeny produced with 55 red plants, 132 pink plants, and 53 white plants. Are these data reasonably consistent with the Mendelian model?

Expected proportion: 1:2:1

Measure: ***Red: 55 | Pink: 132 | White: 53 (Total: 240)***

- **Set up Hypothesis:**

① Hypothesis

H_0: The data are consistent with the ratio 1:2:1

H_a: The data are inconsistent with the model

An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

Hands-On

Inherited Flower Colors

- **Hypothesis**:
  - H_0: The data are consistent with the ratio 1:2:1
  - H_a: The data are inconsistent with the model

② Assumption

General form:

- *chisq.test(table, p, correct = T/F)*

Category	O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
Red	55	60	0.42
Pink	132	120	1.20
White	53	60	0.82
Total	240	240	2.44

- **Practice:**
  - *flower=c(55,132,53)*
  - *chisq.test(flower, p=c(.25,.5,.25), correct=F)*

③ Testing

```
        Chi-squared test for given probabilities

data:  flower
X-squared = 2.4333, df = 2, p-value = 0.2962
```

➡ p = 0.296 > (α = 0.05), Accept H_0

Effect Size of Chi-square Test

④ Effect Size

- Effect size of Chi-square test: use Phi coefficient

Phi coefficient $$\varphi = \sqrt{\frac{\chi^2}{n}}$$

n = the number of observations.

- If contingency table is larger than 2x2: use Cramer's V

Cramer's $$V = \sqrt{\frac{\chi^2}{n \cdot df^*}}$$

$$df^* = \min(r - 1, c - 1)$$

= degree of freedom for the smaller side of the contingency table

Effect Size of Chi-square Test

Phi coefficient $\varphi = \sqrt{\frac{\chi^2}{n}}$

Cramer's $V = \sqrt{\frac{\chi^2}{n \cdot df^*}}$

*df**	*small*	*medium*	*large*
1	.10	.30	.50
2	.07	.21	.35
3	.06	.17	.29
4	.05	.15	.25
5	.04	.13	.22

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

http://www.real-statistics.com/chi-square-and-f-distributions/effect-size-chi-square/

Hands-On

Inherited Flower Colors

④ Effect Size

```
        Chi-squared test for given probabilities

data:  flower
X-squared = 2.4333, df = 2, p-value = 0.2962
```

➡ p = 0.296 > (α = 0.05), Accept H_0

ϕ = sqrt(2.433 / 240) = 0.1

⑤ Decision

- **Reporting decision:**
  The sampled flower colour are reasonably consistent with the Mendelian model (p=0.296).

- **What if we multiply the sample size by 10?**
- **What if we apply Yate's continuity correction?**

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

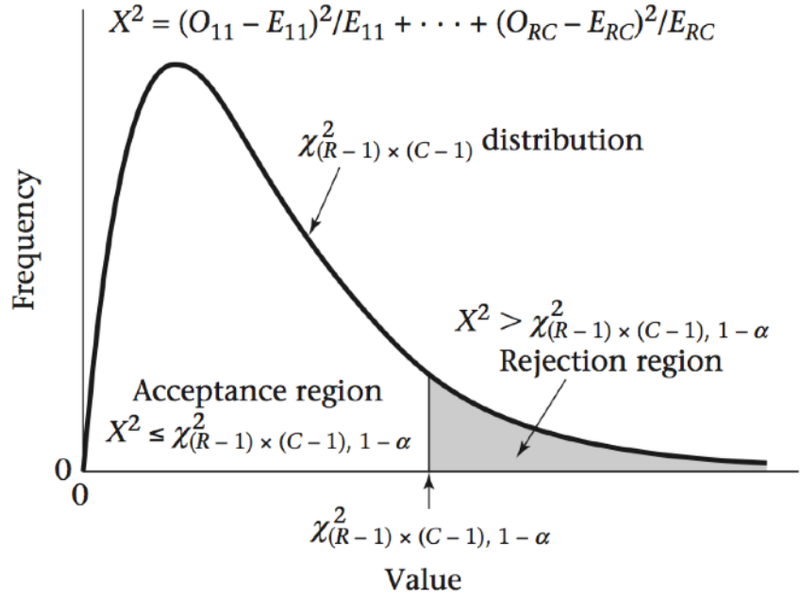
Chi-Square Test & Contingency Table (Test of Independence)

- **Assumptions**
  - An independent random sample of size **n** is drawn from 2 categorical variables. One with **r** categories and the other with **k** categories.
    - not for *repeated-measure* design (use McNemar test instead).
  - The population can be divided into a set of k mutually exclusive categories.
  - The expected frequencies for each category must be specified.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\mathrm{df} = (r-1) \times (k-1).$$

- **Hypotheses**
  - H_0: row and column variables are independent.
  - H_A: row and column variables are associated.



DEMO

Hair & Eye Color

Test of Independence

- Two contingencies for **hair** color: ‘fair’ and ‘dark’
- Two contingencies for **eye** color: ‘blue’ and ‘brown’
- **Are the fair hair and blue eyes correlated with each other?**

	Blue eyes	*Brown eyes*	Row totals
Fair hair	***38***	***11***	49
Dark hair	***14***	***51***	65
Column totals	52	62	114

- **Hypotheses**
  - H_0: row and column variables are independent.
  - H_A: row and column variables are associated.

DEMO

Hair & Eye Color

Test of Independence

- Two contingencies for **hair** color: ‘fair’ and ‘dark’
- Two contingencies for **eye** color: ‘blue’ and ‘brown’
- Are the hair and eye colors dependent to each other?

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Real Observations

	Blue eyes	*Brown eyes*
Fair hair	***38***	***11***
dark hair	***14***	***51***

Expected Frequencies

	Blue eyes	Brown eyes	Row totals
Fair hair	22.35	26.65	49
dark hair	29.65	35.35	65
Column totals	52	62	114

DEMO

Hair & Eye Color

Test of Independence

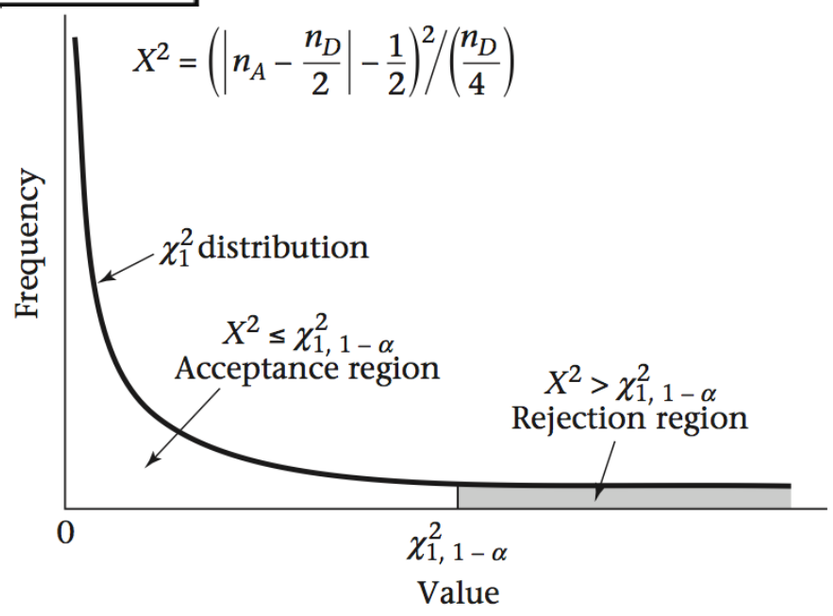
- Two contingencies for **hair** color: ‘fair’ and ‘dark’
- Two contingencies for **eye** color: ‘blue’ and ‘brown’
- Are the hair and eye colors dependent to each other?

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

	O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Fair hair and blue eyes	38	22.35	244.92	10.96
Fair hair and brown eyes	11	26.65	244.92	9.19
Dark hair and blue eyes	14	29.65	244.92	8.26
Dark hair and brown eyes	51	35.35	244.92	6.93

→ $\chi^2 = 35.33$

- **Reporting decision:**
  There is significant positive association between fair hair and blue eyes for this group (p<0.001).



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

DEMO

Age & Breast Cancer

- **Background**: It is speculated that breast cancer in women is caused by events that occur between the age at first menstruation and the age at first childbirth. Researchers want to test whether the age at first childbirth is an risk factor for breast cancer.
- Breast-cancer cases were collected among women in many countries. Controls were chosen from women of comparable age at the same time without breast cancer.

Measure: ***age at first childbirth (30 y/o) & with/without breast cancer***

- **Set up Hypothesis**:

① Hypothesis

H_0: ***Age & BC are independent***

H_a: ***Age is associated with BC***

DEMO

Age & Breast Cancer

- **Hypothesis**:
  - H_0: *Age & BC are independent*
  - H_a: *Age is associated with BC*

- Breast-cancer cases were collected among women in many countries. Controls were chosen from women of comparable age at the same time without breast cancer.

	Age at first birth		
Status	≥30	≤29	Total
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

- **Data import**:
  - *age30 <- c(683, 1498)*
  - *age29 <- c(2547, 8747)*
  - *table <- cbind(age30, age29)*

DEMO

Age & Breast Cancer

- **Hypothesis**:
  - H_0: *Age & BC are independent*
  - H_a: *Age is associated with BC*

Status	Age at first birth ≥30	Age at first birth ≤29	Total
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

- **Practice:**
  - *chisq.test(table, correct=F)*

```
        Pearson's Chi-squared test

data:  table
X-squared = 77.043, df = 1, p-value < 2.2e-16
```

➡ p = 2.2e-16 << (α = 0.05), Reject H_0

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

DEMO

Age & Breast Cancer

④ Effect Size

questionr

```
> odds.ratio(table)
                     OR  2.5 % 97.5 %          p
Fisher's test 1.5658 1.4136 1.7334 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
```

	Age at first birth	
	≥30	≤29
	683	2537
	1498	8747

⑤ Decision

- **Reporting decision:**
  The breast cancer incidence is significantly associated with the age having a first child ($p < 0.001$). After 30 y/o, the odd is 56.6% higher than those having a first child before age 30.

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Fisher's Exact Test

- **Assumptions**
  - An independent random sample of size **n** is drawn from 2 categorical variables.
  - The population can be divided into two mutually exclusive categories.
  - Typically for 2x2 contingency tables when one or more of the expected frequencies less than five (cell < 5). The principle can be extended to a general case of an m x n table.

- **Hypotheses**
  - H_0: A and B factors are **independent**.
  - H_A: A and B factors have association.

2x2 contingency table

a	b	a + b
c	d	c + d
a + c	b + d	n

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!}$$

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

https://en.wikipedia.org/wiki/Fisher%27s_exact_test

DEMO

Tree & Ant

- Does tree type has relationship with the ant colonies?

	Tree A	*Tree B*
With Ants	***6***	***2***
Without Ants	***4***	***8***

H_0: Tree types and ants are independent.

H_A: Tree types and ants have association.

```
Fisher's Exact Test for Count Data

data: x
p-value = 0.1698
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6026805 79.8309210
sample estimates:
odds ratio
  5.430473
```

- **Reporting decision:**
  No significant association between tree types and ant colonies (*p*=.17).

Discussion

1. **Proportions**
   - Binomial test for percentage or rates
2. **Counts**
   - Chi-square test
3. **[R] Hands-on Practices**