✪ Psychological Statistics ✪

Week 06: *Counts, Table & Chi-Square Test*

- Edited by Prof. **Changwei Wu**
- Graduate Institute of Mind, Brain and Consciousness (***GIMBC***), Taipei Medical University

In [58]:
```
### [ Loading the required libraries ]

library("questionr")
library("dplyr")
library("rstatix")
library("effsize")
library("ggpubr")
```

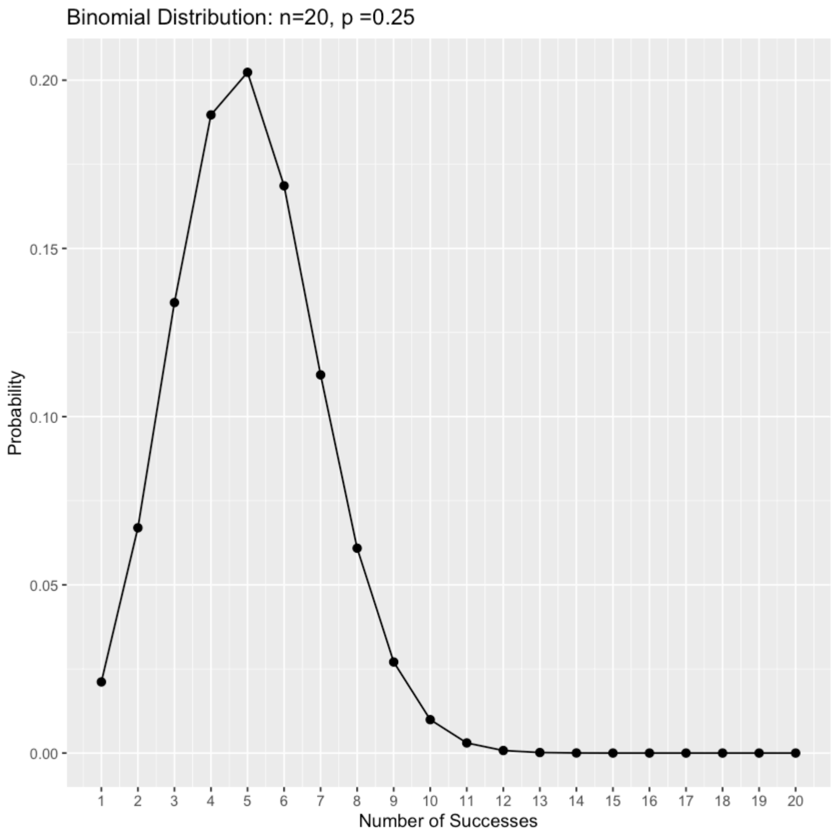
---

(1) Proportions: Bionomial distribution

In [53]:
```r
### [ Distribution plot ] Binomial distribution

n = 20
p = 0.25

binom_dist <- tibble(n_success = 1:n) %>%
  mutate(probability = dbinom(n_success, size=n, prob=p))
#binom_dist
binom_dist %>%
  ggplot(aes(x=n_success, y=probability))+
  geom_line()+
  geom_point(size=2)+
  scale_x_continuous(breaks=1:n)+
  scale_y_continuous(breaks = scales::pretty_breaks(n = 5))+
  labs(x= "Number of Successes", y= "Probability",
       title=paste0("Binomial Distribution: n=",n,", p =",p))
```



In [31]:
```r
### [ Example 3.9 ] Mice muscle destrophy


# → (a) Fewer than 5 will have muscular dystrophy: P(X<5)

pbinom(4,20,0.25) %>% round(3)

# → (b) 5 will have muscular dystrophy: P(X=5)

dbinom(4,20,0.25) %>% round(2)
```

0.415

0.19

✪ Normal approximation (increasing trial number)

prop_test {rstatix} | prop.test {stats}

In [9]:
```
### [ Example 3.18 ] Mice muscle destrophy -- Normal approximation
# → Binomial probability with P = 0.25 and n = 20


# What is the probability of fewer than 15 with muscular dystrophy out of 60?

(binom_orig <- pbinom(14, 60, 0.25) %>% round(3))

# Same question with Normal Approximation [mean = np; var = sqrt(npq)]

(binom_Zdist_noC <- pnorm(14, 60*.25, sqrt(60*.25*.75)) %>% round(3))

(binom_Zdist_wthC <- pnorm(14+0.5, 60*.25, sqrt(60*.25*.75)) %>% round(3))
```

0.451

0.383

0.441

Example 11.2: Special thin growth ring of 1987 (Compared to a predefined proportion)

[Hypothesis] **Majority of trees have the special growth ring.** ***(1-tailed)***

- Null hypothesis H_0: Tree(1987ring) ≤ 0.5
- Alternative hyp. H_1: Tree(1987ring) > 0.5

In [10]:
```
### [ Step.1 ] Load data

(p_1987ring <- 15/20)
```

0.75

In [ ]:
```
### [ Step.2 ] Assumption check

# → No specific assumption to be checked for binomial distribution
```

In [12]:
```
### [ Step.3 ] Binomial Test on the counts of trees

(Tree.test <- binom_test(x=15, n=20, p = 0.5, alternative = "greater"))
```

A rstatix_test: 1 × 6

	n	estimate	conf.low	conf.high	p	p.signif
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	20	0.75	0.5444176	1	0.02069473	*

In [13]:
```
### [ Step.4 ] Effect size
# → Odd's ratio for Proportions

(ODD_ratio <- 0.75/0.5)
```

1.5

~ *Report* ~

- Significant evidence exhibits that the majority of trees (75%) have growth rings of 1987 less than half their usual size ($p < 0.021$).

> ## Example 11.6: Death rate after stenting (Comparing 2 proportions)
>
> ### *[Hypothesis]* Stenting surgery saves lives. *(2-tailed)*
>
> - Null hypothesis H_0 : Death_rate(with stents) = Death_rate(without stents)
> - Alternative hyp. H_1 : Death_rate(without stents) ≠ Death_rate(without stents)

In [35]:
```
### [ Step.1 ] Load data

stent.data <- as.table(rbind(c(171, 179), c(1082, 1084)))
dimnames(stent.data) <- list(
    case=c("Death", "Total"),
    group=c("Stent", "No Stent"))
stent.data
```

```
       group
case    Stent No Stent
  Death   171      179
  Total  1082     1084
```

In [43]:
```
### [ Step.1 ] Load data

stent.data <- as.table(rbind(c(171, 1082), c(179, 1084)))
dimnames(stent.data) <- list(
    group=c("Stent", "No Stent"),
    case=c("Death", "Total"))
stent.data
```

```
          case
group      Death Total
  Stent      171  1082
  No Stent   179  1084
```

In [29]:
```
### [ Step.2 ] Assumption check

# → Independent & Mutually exclusive for binomial distribution (not testable)
```

In [48]:
```
### [ Step.3 ] Using prop_test | prop.test for compring 2 proportions

(Death.Stent <- prop_test(stent.data, alternative = "less", correct=F))

prop.test(stent.data, alternative = "less", correct=F)
```

A rstatix_test: 1 × 5

	n	statistic	df	p	p.signif
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	2516	0.1044111	1	0.373	ns

```
	2-sample test for equality of proportions without continuity
	correction

data:  stent.data
X-squared = 0.14496, df = 1, p-value = 0.3517
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01744079
sample estimates:
   prop 1    prop 2 
0.1364725 0.1417260 
```

In [51]:
```
### [ Step.4 ] Effect size
# → Phi
cramer_v(stent.data, correct=F)
```

0.00759048560659339

~ *Report* ~

- There is no evidence for a significant reduction in death rate with the implantation of stents in patients after heart attack (p = 0.35).

---

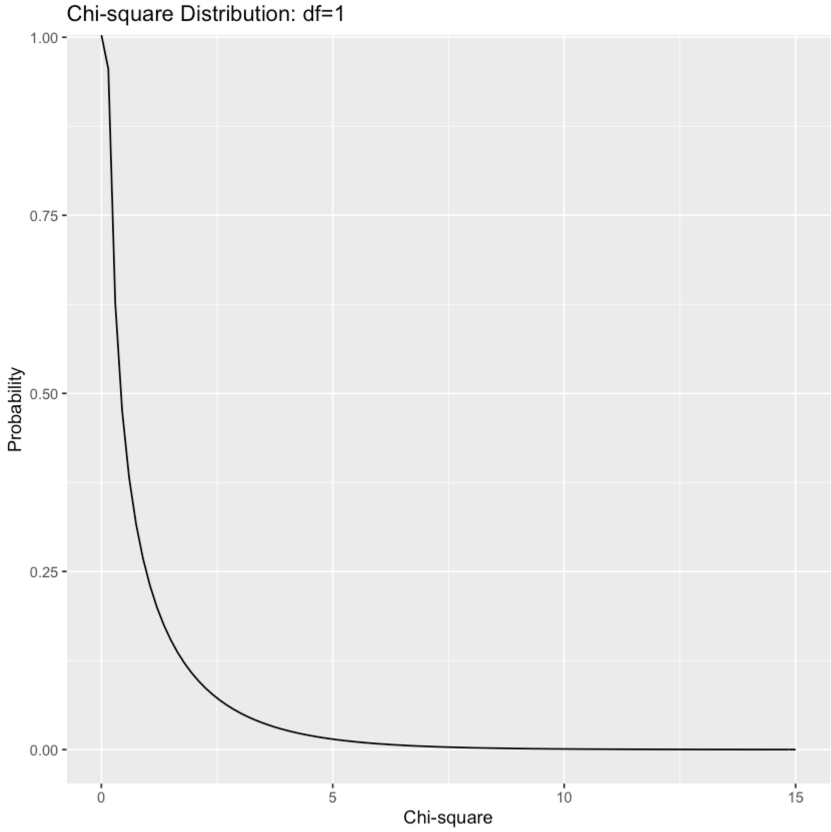
(2) Data counts by Chi-Square Test

In [57]:
```
### [ Distribution plot ] Chi-square distribution

df = 1
xvec <- seq(0,15,length=101)

pvec <- dchisq(xvec,df)

ggplot(data.frame(x = c(0, 15)), aes(x = x)) +
     stat_function(fun = dchisq, args = list(df))+
  labs(x= "Chi-square", y= "Probability",
       title=paste0("Chi-square Distribution: df=",df))
```



Data count: freq_table {rstatix} | xtabs {stats}

In [22]:
```
### 2.1 [ Frequency count (1)]
data("ToothGrowth")

ToothGrowth %>% freq_table(supp)

xtabs(~supp, ToothGrowth)
```

A tibble: 2 × 3

supp	n	prop
<fct>	<int>	<dbl>
OJ	30	50
VC	30	50

```
supp
OJ VC
30 30
```

In [21]:
```
### 2.2 [ Frequency count (2)]

ToothGrowth %>% freq_table(supp, dose)

xtabs(~supp + dose, ToothGrowth)
```

A tibble: 2 × 3

supp	n	prop
<fct>	<int>	<dbl>
OJ	30	50
VC	30	50

```
supp
OJ VC
30 30
```

Chi-Square Test: chisq_test {rstatix} | chi.test {stats}

> ### (1) Goodness of Fit (compared with reference proportions)
>
> ### Example: Flowers
>
> ### *[Hypothesis]* Are the flower colors equally common? *(2-tailed)*
>
> - Null hypothesis ***H_0* :** flower colors are equally common
> - Alternative hyp. ***H_1* :** flower colors are NOT equally common

In [67]:
```
### [ Step.1 ] Load data

flower <- c(red = 55, pink = 132, white = 53)
```

In [68]:
```
### [ Step.3 ] Chi-square test

chisq_test(flower, p = c(0.25, 0.5, 0.25))
```

A rstatix_test: 1 × 6

	n	statistic	p	df	method	p.signif
	<int>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
1	3	2.433333	0.296	2	Chi-square test	ns

In [69]:
```
### [ Step.4 ] Effect size

pairwise_chisq_test_against_p(flower, , p = c(0.25, 0.5, 0.25))
```

A rstatix_test: 3 × 9

	group	observed	expected	n	statistic	p	df	p.adj	p.adj.signif
	<chr>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	red	55	60	2	0.5555556	0.456	1	0.594	ns
2	pink	132	120	2	2.4000000	0.121	1	0.363	ns
3	white	53	60	2	1.0888889	0.297	1	0.594	ns

~ *Report* ~

- The sampled flower colour are reasonably consistent with the Mendelian model (*p* = .296).

(2) Homogeneity of proportions (between groups)

Example: Survivors of Titanic

[Hypothesis] Survival rate of different classes are equal. *(2-tailed)*

- Null hypothesis H_0 **:** Survival rate of 4 groups are similar.
- Alternative hyp. H_1 **:** Survival rate of 4 groups are different.

In [3]:
```
### [ Step.1 ] Load data

titanic.surv <- as.table(rbind(
  c(203, 118, 178, 212),
  c(122, 167, 528, 673)
))

dimnames(titanic.surv) <- list(
  Survived = c("Yes", "No"),
  Class = c("1st", "2nd", "3rd", "Crew")
)
titanic.surv
```

```
        Class
Survived 1st 2nd 3rd Crew
     Yes 203 118 178  212
     No  122 167 528  673
```

In [4]:
```
### [ Step.3 ] Chi-square test

chisq_test(titanic.surv)
```

A rstatix_test: 1 × 6

	n	statistic	p	df	method	p.signif
	<dbl>	<dbl>	<dbl>	<int>	<chr>	<chr>
1	2201	190.4011	5e-41	3	Chi-square test	****

In [6]:
```
### [ Step.4 ] Effect size

cramer_v(titanic.surv)

pairwise_prop_test(titanic.surv)
```

0.294120103005126

A rstatix_test: 6 × 5

	group1	group2	p	p.adj	p.adj.signif
	<chr>	<chr>	<dbl>	<dbl>	<chr>
1	1st	2nd	3.13e-07	9.38e-07	****
2	1st	3rd	2.55e-30	1.27e-29	****
3	2nd	3rd	6.90e-07	1.38e-06	****
4	1st	Crew	1.62e-35	9.73e-35	****
5	2nd	Crew	1.94e-08	7.75e-08	****
6	3rd	Crew	6.03e-01	6.03e-01	ns

~ *Report* ~

- The survival rates are different (p < .001) between different classes in Titanic.

---

(3) Test of Independence (between factors)

Example: Color of eyes and color of hair

[Hypothesis] Brown eyes leads to the dark hair. *(2-tailed)*

- Null hypothesis H_0 **:** Eye color and hair color are independent.
- Alternative hyp. H_1 **:** Eye color and hair color are correlated.

In [70]:
```
### [ Step.1 ] Load data
colors <- as.table(rbind(
  c(38, 11),
  c(14, 51)
))
dimnames(colors) <- list(
  Hair = c("Fair", "Dark"),
  Eyes = c("Blue", "Brown")
)
colors
```

```
      Eyes
Hair   Blue Brown
  Fair   38    11
  Dark   14    51
```

In [71]:
```
### [ Step.3 ] Chi-square test

chisq_test(colors)
```

A rstatix_test: 1 × 6

	n	statistic	p	df	method	p.signif
	<dbl>	<dbl>	<dbl>	<int>	<chr>	<chr>
1	114	33.11197	8.7e-09	1	Chi-square test	****

~ *Report* ~

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

---

(3) CrossTable {gmodels}

> ### Example: Age vs. Breast cancer
>
> ### *[Hypothesis]* The age at first childbirth is an risk factor for breast cancer. *(2-tailed)*
>
> - Null hypothesis H_0 **:** Birth-giving age and breast cancer are 2 independent factors.
> - Alternative hyp. H_1 **:** Birth-giving age and breast cancer has certain relationship with each other.

In [72]:
```
### [ Step.1 ] Load data

Age_BC <- as.table(rbind(
  c(683, 2537),
  c(1498, 8747)
))
dimnames(Age_BC) <- list(
  Status = c("BCcase", "Control"),
  Eyes = c("age≥30", "age<30")
)
Age_BC
```

```
         Eyes
Status    age≥30 age<30
  BCcase     683   2537
  Control   1498   8747
```

In [77]:

```
### [ Step.3 ] CrossTable

#CrossTable(Age_BC)
CrossTable(Age_BC, fisher = TRUE, chisq = TRUE, expected = TRUE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  13465


             | Eyes
      Status |    age≥30 |    age<30 | Row Total |
-------------|-----------|-----------|-----------|
      BCcase |       683 |      2537 |      3220 |
             |   521.561 |  2698.439 |           |
             |    49.970 |     9.658 |           |
             |     0.212 |     0.788 |     0.239 |
             |     0.313 |     0.225 |           |
             |     0.051 |     0.188 |           |
-------------|-----------|-----------|-----------|
     Control |      1498 |      8747 |     10245 |
             |  1659.439 |  8585.561 |           |
             |    15.706 |     3.036 |           |
             |     0.146 |     0.854 |     0.761 |
             |     0.687 |     0.775 |           |
             |     0.111 |     0.650 |           |
-------------|-----------|-----------|-----------|
Column Total |      2181 |     11284 |     13465 |
             |     0.162 |     0.838 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  78.36984     d.f. =  1     p =  8.544684e-19

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  77.88515     d.f. =  1     p =  1.092096e-18


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  1.571925

Alternative hypothesis: true odds ratio is not equal to 1
p =  5.873474e-18
95% confidence interval:  1.419073 1.740189

Alternative hypothesis: true odds ratio is less than 1
p =  1
95% confidence interval:  0 1.712541

Alternative hypothesis: true odds ratio is greater than 1
p =  3.526441e-18
95% confidence interval:  1.442384 Inf
```

In [79]:
```
### [ Step.4 ] Effect size
# → Function: odds.ratio() {questionr}

odds.ratio(Age_BC)
```

```
Registered S3 method overwritten by 'DescTools':
  method         from
  reorder.factor gdata
```

1.57198193044674

A odds.ratio: 1 × 4

	OR	2.5 %	97.5 %	p
	<dbl>	<dbl>	<dbl>	<dbl>
Fisher's test	1.571925	1.419073	1.740189	5.873474e-18

~ *Report* ~

- The breast cancer incidence is significantly associated with having a first child after age 30 ($p < 0.001$). Their odd is 56.6% higher than those having a first child before age 30.