2019-2

Introduction to Biostatistics - Lecture 3: Statistical Inference for Proportions

Jonggyu Baek
University of Massachusetts Medical School

University of Massachusetts Medical School
UMASS
Department of Quantitative Health Sciences

Introduction to Biostatistics

2/29/2019

Jonggyu Baek, PhD

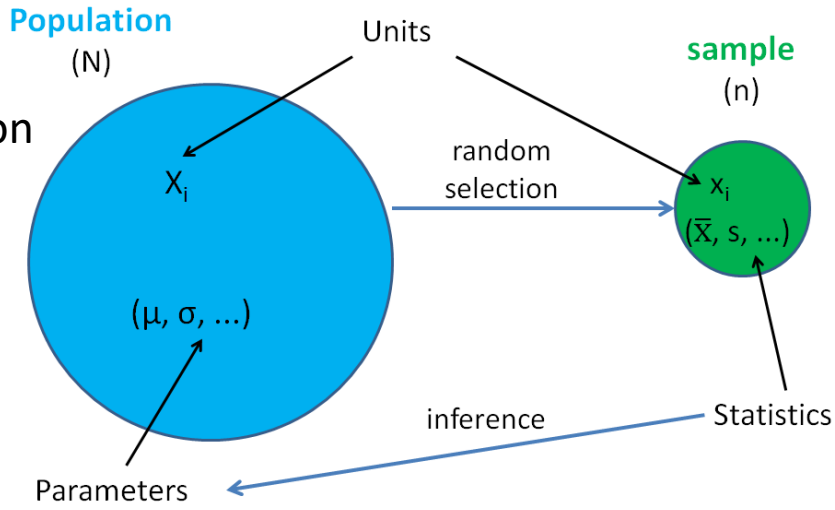
Lecture 3:

Statistical Inference for proportions

Statistical Inference

Two broad areas of statistical inference:

- **Estimation:** Use sample statistics to estimate the unknown population parameter.
  - **Point Estimate:** the best single value to describe the unknown parameter.
  - **Standard Error (SE):** standard deviation of the sample statistic. Indicates how precise is the point estimate.
  - **Confidence Interval (CI):** the range with the most probable values for the unknown parameter with a (1-α)% level of confidence.



- **Hypothesis Testing:** Test a specific statement (assumption) about the unknown parameter.

Statistical Inference for proportions

Suppose X: discrete (binary) variable with:

$$X=\begin{cases}1, & \text{event A with probability p}\\ 0, & \text{otherwise with probability } 1-\text{p}\end{cases}$$

We are interested in estimating the probability p of the event A in a population of size N:

P(event A) =

Statistical Inference for proportions

Suppose X: discrete (binary) variable with:

$$X=\begin{cases}1, & \text{event A with probability p}\\ 0, & \text{otherwise with probability } 1-\text{p}\end{cases}$$

We are interested in estimating the probability p of the event A in a population of size N:

$$\text{P(event A)} = \frac{\text{\# of favorable outcomes}}{\text{sample space}} = \frac{\text{\# of successes}}{\text{Total \# of units in the population}} \quad (1)$$

Statistical Inference for proportions

Suppose X: discrete (binary) variable with:

$$X=\begin{cases}1, & \text{event A with probability p}\\ 0, & \text{otherwise with probability } 1-\text{p}\end{cases}$$

We are interested in estimating the probability p of the event A in a population of size N:

$$\text{P(event A)} = \frac{\text{\# of favorable outcomes}}{\text{sample space}} = \frac{\text{\# of successes}}{\text{Total \# of units in the population}} \quad (1)$$

Suppose Y = # of successes = $\sum_{i=1}^{N} X_i$. Then Y ~ Binomial(N, p)

Statistical Inference for proportions

Suppose X: discrete (binary) variable with:

$$X=\begin{cases}1, & \text{event A with probability p}\\ 0, & \text{otherwise with probability } 1-\text{p}\end{cases}$$

We are interested in estimating the probability p of the event A in a population of size N:

$$\text{P(event A)} = \frac{\text{\# of favorable outcomes}}{\text{sample space}} = \frac{\text{\# of successes}}{\text{Total \# of units in the population}} \quad (1)$$

Suppose Y = # of successes = $\sum_{i=1}^{N} X_i$. Then Y ~ Binomial(N, p)

$$\text{P(event A)} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Statistical Inference for proportions

Suppose X: discrete (binary) variable with:

$$X=\begin{cases}1, & \text{event A with probability p}\\ 0, & \text{otherwise with probability } 1-p\end{cases}$$

We are interested in estimating the probability p of the event A in a population of size N:

$$P(\text{event A}) = \frac{\#\text{ of favorable outcomes}}{\text{sample space}} = \frac{\#\text{ of successes}}{\text{Total \# of units in the population}} \quad (1)$$

Suppose Y = # of successes = $\sum_{i=1}^{N} X_i$. Then Y ~ Binomial(N, p)

What is this?

$$P(\text{event A}) = \frac{\sum_{i=1}^{N} X_i}{N}$$

Statistical Inference for proportions

Suppose X: discrete (binary) variable with:

$$X=\begin{cases}1, & \text{event A with probability p}\\ 0, & \text{otherwise with probability } 1-p\end{cases}$$

We are interested in estimating the probability p of the event A in a population of size N:

$$P(\text{event A}) = \frac{\#\text{ of favorable outcomes}}{\text{sample space}} = \frac{\#\text{ of successes}}{\text{Total \# of units in the population}} \quad (1)$$

Suppose Y = # of successes = $\sum_{i=1}^{N} X_i$. Then Y ~ Binomial(N, p)

$$P(\text{event A}) = \frac{\sum_{i=1}^{N} X_i}{N} = p$$

What is this?

- A proportion
- A population mean

Hence, all **the statistical inference procedures** we learned about the **means** also apply for **proportions**.

Statistical Inference for proportions

- **Case 1:** single population (**one-sample**)
- **Case 2:** two-**independent** populations (**two-samples**)
- **Case 3:** two-**dependent** populations (**paired** or **matched** samples)

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p}(\mathbf{1}-\mathbf{p})}$.

Estimation

- **Point Estimates:**
  - of p: $\bar{x} = \hat{p}$
  - of σ: $s = \sqrt{\hat{p}(1-\hat{p})}$
  - precision of $\bar{x}$: standard error (s.e.) of $\hat{p} \rightarrow \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **(1-α)% CI:**

$$[\ \hat{p} - Z_{1-\alpha/2} \cdot \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\ ,\ \hat{p} + Z_{1-\alpha/2} \cdot \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)]$$

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p}(\mathbf{1}-\mathbf{p})}$.

- **Point Estimates:**

```
library(sjmisc)
frq(dat1$stroke)
```

```
> frq(dat1$stroke)

# x <integer>
# total N=11627  valid N=11627  mean=0.09  sd=0.29

  val   frq raw.prc valid.prc cum.prc
    0 10566   90.87     90.87   90.87
    1  1061    9.13      9.13  100.00
 <NA>     0    0.00        NA      NA
```

- If the binary variable is code as ["1"=yes, "0"=no] we can also calculate the mean:

```
## the estimated proportion of stroke ##
mean(dat1$stroke)
```

```
> mean(dat1$stroke)
[1] 0.09125312
```

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p}(\mathbf{1}-\mathbf{p})}$.

- **Point Estimates:**

```
> frq(dat1$stroke)

# x <integer>
# total N=11627  valid N=11627  mean=0.09  sd=0.29

  val   frq raw.prc valid.prc cum.prc
    0 10566   90.87     90.87   90.87
    1  1061    9.13      9.13  100.00
 <NA>     0    0.00        NA      NA
```

- If the binary variable is code as ["1"=yes, "0"=no] we can also calculate the mean:

```
## the estimated proportion of stroke ##
mean(dat1$stroke)

> mean(dat1$stroke)
[1] 0.09125312
```

$\hat{\mathbf{p}}$

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p(1-p)}}$.

- **What about σ=**$\sqrt{\mathbf{p(1-p)}}$ **?:**

```
## the standard deviation of stroke ##
p = mean(dat1$stroke)
std.dev = sqrt(p*(1-p))
p
std.dev
```

```
> ## the standard deviation of stroke ##
> p = mean(dat1$stroke)
> std.dev = sqrt(p*(1-p))
> p
[1] 0.09125312
> std.dev
[1] 0.2879687
```

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p(1-p)}}$.

Hypothesis Testing

- Null hypothesis (H_0): $p=p_0$
- Alternative hypothesis (H_1):
  - $p \neq p_0$ (two-sided test), or
  - $p < p_0$ (one-sided test), or
  - $p > p_0$ (one-sided test)

- Test statistic: $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{H_0}{\sim} N(0, 1)$

Why?

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p(1-p)}}$.

Hypothesis Testing

- Null hypothesis (H_0): $p=p_0$
- Alternative hypothesis (H_1):
  - $p \neq p_0$ (two-sided test), or
  - $p < p_0$ (one-sided test), or
  - $p > p_0$ (one-sided test)

- Test statistic:

$$Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{H_0}{\sim} N(0, 1)$$

Why?

From the CLT:

$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

i.e.,

$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p(1-p)}}$.

Hypothesis Testing

- Test statistic: $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{H_0}{\sim} N(0, 1)$
- Decision Rules by H_1: Testing H_0: $p=p_0$ vs :

H_1	Reject H_0 if:
$p \neq p_0$	$Z_0 < Z_{\alpha/2}$ or $Z_0 > Z_{1-\alpha/2}$
$p < p_0$	$Z_0 < Z_{\alpha}$
$p > p_0$	$Z_0 > Z_{1-\alpha}$

One Sample

- **Case 1:** single population (one-sample)

Example (FHS):

- Calculate 95% CI for the proportion of strokes in the population
- Test whether this proportion is not different from 0.12=12%

```
library(sjmisc)
frq(dat1$stroke)
```

```
> frq(dat1$stroke)

# x <integer>
# total N=11627  valid N=11627  mean=0.09  sd=0.29

  val   frq raw.prc valid.prc cum.prc
    0 10566   90.87     90.87   90.87
    1  1061    9.13      9.13  100.00
 <NA>     0    0.00        NA      NA
```

One Sample

Case 1: single population (one-sample)

Suppose X='**stroke**' from a population with mean **μ=p** and standard deviation **σ=**$\sqrt{\mathbf{p}(\mathbf{1}-\mathbf{p})}$.

Hypothesis Testing

- Testing H_0: $p=p_0=12\%=0.12$

```
prop.test(x=1061, n=11627, p = 0.12, alternative="two.sided")

> prop.test(x=1061, n=11627, p = 0.12, alternative="two.sided")

        1-sample proportions test with continuity correction

data:  1061 out of 11627, null probability 0.12
X-squared = 90.716, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.12
95 percent confidence interval:
 0.08611105 0.09666741
sample estimates:
         p
0.09125312
```

One Sample

- **Case 1:** single population (one-sample)

  X: discrete (**binary**) variable (e.g., '**stroke**')

  Statistical Inference about:

  p (Proportion of strokes in the population)

ESTIMATION		HYPOTHESIS TESTING (H_0: $\rho=\rho_0$)	
Point Estimate	$\hat{p}$	Test Statistic	$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
Standard Error	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Decision rules Reject H_0 against H_1:	
(1-α)% CI	$\hat{p} \pm Z_{1-\alpha/2} \cdot \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$	$\mu \neq \mu_0$ $\mu < \mu_0$ $\mu > \mu_0$	$Z_0 < Z_{\alpha/2}$ or $Z_0 > Z_{1-\alpha/2}$ $Z_0 < Z_\alpha$ $Z_0 > Z_{1-\alpha}$

Two Independent Samples

- **Case 2:** two-independent populations (two-samples)

Suppose **Y**='**stroke**' and **X**='**prevchd**'

(both **binary** variables).

There are two-independent populations, one with coronary heart disease (chd) and the other without chd. We want to compare proportions of strokes between those two populations.

- $\mathbf{p_1}$ is the proportion of strokes in the population **with CHD**.
- $\mathbf{p_2}$ is the proportion of strokes in the population **without CHD**.

Two Independent Samples

- **Case 2:** two-independent populations (two-samples)

  Y='stroke' and **X='prevchd'** (both **binary** variables)

  Statistical Inference about:

  p_1-p_2

  (compare proportions of strokes between the two populations)

ESTIMATION		HYPOTHESIS TESTING (H_0: p_1-p_2=0)	
Point Estimate	$\hat{p}_1$-$\hat{p}_2$	Test Statistic	$Z = \frac{\hat{p}_1 - \hat{p}_2}{s.e.}$
Standard Error	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Decision rules Reject H_0 against H_1:	(for 'small' n_1, n_2 use the Binomial distribution – exact test)
(1-α)% CI	$\hat{p}_1 - \hat{p}_2 \pm Z_{1-\alpha/2} \cdot s.e.$	$\mu \neq \mu_0$ $\mu < \mu_0$ $\mu > \mu_0$	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2}$ $Z<Z_{\alpha}$ $Z>Z_{1-\alpha}$

Two Independent Samples

- **Case 2:** two-independent populations (two-samples)

Example (FHS):

  - Test whether the probability of stroke (**binary**) is equal between people with and without CHD (**binary**).

```
> ## two sample proportion test ##
> tab2 = table(dat1$stroke, dat1$prevchd, deparse.level = 2)
> tab2
           dat1$prevchd
dat1$stroke    0    1
          0 9887  679
          1  898  163
> prop.table(tab2, margin=2)
           dat1$prevchd
dat1$stroke          0          1
          0 0.91673621 0.80641330
          1 0.08326379 0.19358670
> prop.test(x = c(898, 163), n=c(10785, 842)) ## or prop.test(x = c(898, 163), n=table(dat1$prevchd))

        2-sample test for equality of proportions with continuity correction

data:  c(898, 163) out of c(10785, 842)
X-squared = 113.31, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.13815524 -0.08249057
sample estimates:
    prop 1     prop 2
0.08326379 0.19358670
```

Two Dependent Samples

- **Case 3:** two-dependent populations (two-samples)

Suppose **Y1**='smoke at period 1' and **Y2**='smoke at period 2'
(both **binary** variables).

Suppose that we collect info at baseline (before) and 6 years after some anti-hypertensive treatment. We want to compare the proportions of smoke status between those two time points.

- $\mathbf{p_1}$ is the proportion of smoking status at period 1.
- $\mathbf{p_2}$ is the proportion of smoking status at period 2.

We apply the **McNemar's Test**.

This test ONLY depends on the number of discordant pairs.

Long format to short format

```
## two dependent sample proportion test ##
dat2 = dat1[,c("randid", "period", "cursmoke")] ## to only select those three variables ##
head(dat2)

library(tidyr) # to transform a long data format to a short data format

dat_short = spread(dat2, period, value=cursmoke)
head(dat_short)
```

```
> ## two dependent sample proportion test ##
> dat2 = dat1[,c("randid", "period", "cursmoke")] ## to only select those three variables ##
> head(dat2)
  randid period cursmoke
1   2448      1        0
2   2448      3        0
3   6238      1        0
4   6238      2        0
5   6238      3        0
6   9428      1        1
>
> library(tidyr) # to transform a long data format to a short data format
>
> dat_short = spread(dat2, period, value=cursmoke)
> head(dat_short)
  randid 1  2  3
1   2448 0 NA  0
2   6238 0  0  0
3   9428 1  1 NA
4  10552 1  1 NA
5  11252 1  1  1
6  11263 0  0  0
```

Long format to short format

- **To change variable names in R**

```
## to change variable names ##
names(dat_short) = c("randid", "cursmoke1", "cursmoke2", "cursmoke3")
head(dat_short)
```

```
> ## to change variable names ##
> names(dat_short) = c("randid", "cursmoke1", "cursmoke2", "cursmoke3")
> head(dat_short)
  randid cursmoke1 cursmoke2 cursmoke3
1   2448         0        NA         0
2   6238         0         0         0
3   9428         1         1        NA
4  10552         1         1        NA
5  11252         1         1         1
6  11263         0         0         0
```

- **Proportion of smoke at period 1 and period 2**

```
mean(dat_short$cursmoke1)
mean(dat_short$cursmoke2, na.rm=TRUE)
```

```
> mean(dat_short$cursmoke1)
[1] 0.4918809
> mean(dat_short$cursmoke2, na.rm=TRUE)
[1] 0.4394402
```

Two Dependent Samples

- **Case 3:** two-dependent populations (two-samples)

Example (FHS):

- Test whether the probability of stroke is equal between the first two periods.

```
## to test two dependent binary variables ##
tab1 = table(dat_short$cursmoke1, dat_short$cursmoke2, deparse.level = 2)
tab1

mcnemar.test(tab1)
```

```
> tab1 = table(dat_short$cursmoke1, dat_short$cursmoke2, deparse.level = 2)
> tab1
                   dat_short$cursmoke2
dat_short$cursmoke1    0    1
                  0 1898  131
                  1  305 1596
>
> mcnemar.test(tab1)

        McNemar's Chi-squared test with continuity correction

data:  tab1
McNemar's chi-squared = 68.644, df = 1, p-value < 2.2e-16
```

Statistical Inference for proportions

- So far we have talked about X discrete, **binary** variables.

- What if X is NOT binary, i.e., it has more than two (>2) groups ?

Statistical Inference for proportions

- So far we have talked about Y, X discrete, **binary** variables.

- What if Y, X are **NOT binary**, i.e., they have more than two (>2) groups ?

We apply **Chi-square tests**.

Chi-square test can also be used when there are two groups as well☺

Statistical Inference for proportions

Comparing Risks in ≥ 2 populations

Statistical Inference for proportions
Fisher's Exact test

- **Test of Independence or Homogeneity:**
  - To test hypotheses concerning the risks in two-populations

  H_0: the two population risks (R_1 and R_2) are similar

  H_1: R_1 and R_2 are different

Statistical Inference for proportions

- There are several ways to express difference in risks.
- **Effect Measures:** measures of difference in risks
  - Risk difference: $\widehat{RD} = \hat{p}_1 - \hat{p}_0$
  - Relative risk difference: $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0}$
  - Odds ratio: $\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}$

where: $\hat{p}_1 = \frac{a}{a+b}$ and $\hat{p}_0 = \frac{c}{c+d}$

- No intuitive interpretation of the **OR**.
- For small $\hat{p}$ (**rare events**) **OR** is a very good estimate of the **RR**.

Comparison Group	Outcome		Total
	1	0	
1	a	b	**a+b**
0	c	d	**c+d**
Total	**a+c**	**b+d**	**N=a+b+c+d**

Confidence Intervals for Effect Measures

- Risk difference: $\widehat{RD} = \hat{p}_1 - \hat{p}_0$

$$\widehat{RD} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$$

- Relative risk difference: $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0}$

$$\exp\left(\ln(\widehat{RR}) \pm Z_{1-\alpha/2}\sqrt{\frac{d/c}{n_0} + \frac{(b/a)}{n_1}}\right)$$

- Odds ratio: $\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}$

$$\exp\left(\ln(\widehat{OR}) \pm Z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

where:

$$\hat{p}_1 = \frac{a}{a+b}$$

$$\hat{p}_0 = \frac{c}{c+d}$$

Comparison Group	Outcome		Total
	1	0	
1	a	b	n_1=a+b
0	c	d	n_0=c+d
Total	**a+c**	**b+d**	**N=a+b+c+d**

Statistical Inference for proportions
Chi-square Tests

- **Test of independence or homogeneity:**

  Example: compare the risk of 'stroke' between people receiving (R_1) and not-receiving (R_0) anti-hypertensive treatment.

  We can express the **null hypothesis** as:

$$H_0: \text{RD}=0 \Rightarrow R_1 - R_2 = 0$$

$$H_1: R_1 - R_2 \neq 0$$

or

$$H_0: \text{RR}=1 \Rightarrow R_1 / R_2 = 1$$

$$H_1: R_1 / R_2 \neq 1$$

or

$$H_0: \text{OR}=1 \Rightarrow \text{Odds}_1 / \text{Odds}_2 = 1$$

$$H_1: \text{Odds}_1 / \text{Odds}_2 \neq 1$$

Statistical Inference for proportions

- H_0: the risk of having the stroke is the same between the two people receiving and not-receiving antihypertensive treatment.
- or
- H_0: there is NO difference in the risk of having the stroke between the two groups

```
## 2x2 table of stroke and bpmeds ##
tab1 = table(dat1$stroke, dat1$bpmeds)
prop.table(tab1, margin=2)

## chisquare test ##
chisq.test(tab1)
```

```
> ## 2x2 table of stroke and bpmeds ##
> tab1 = table(dat1$stroke, dat1$bpmeds)
> prop.table(tab1, margin=2)

             0         1
  0 0.92051536 0.78601695
  1 0.07948464 0.21398305
> ## chisquare test ##
> chisq.test(tab1)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab1
X-squared = 187.17, df = 1, p-value < 2.2e-16
```

Statistical Inference for proportions
Chi-square Tests

- **Test of Independence:**
  - Two or more populations each of which can be split in q≥2 groups (G_1, G_2, …, G_q) according to some characteristic

    e.g., 'bmi' categories by 'sysbp' groups

where, $$\text{bmi} = \begin{cases} \text{underweight } (< 18.5) \\ \text{normal } (18.5 - 25) \\ \text{overweight } (25 - 30) \\ \text{obese } (> 30) \end{cases}$$ and $$\text{sysbp} = \begin{cases} \text{normal } (< 120) \\ \text{moderate } (120 - 140) \\ \text{high } (> 140) \end{cases}$$

  - We want to test the hypothesis that the distribution of bmi categories is the same for each sysbp group, or (in other words) that 'bmi' is independent of 'sysbp'

H_0: the two characteristics (X_1 and X_2) are independent

H_1: H_0 is false

Statistical Inference for proportions
Chi-square Tests

- **Categorizing a continuous variable in R**

```
### more than 2 categories categorical bmi vs. categorical sysbp ###
dat1$cat_bmi = NULL
dat1$cat_bmi[dat1$bmi<18.5] = "underweight"
dat1$cat_bmi[dat1$bmi>=18.5 & dat1$bmi < 25] = "normal"
dat1$cat_bmi[dat1$bmi>=25 & dat1$bmi < 30] = "overweight"
dat1$cat_bmi[dat1$bmi>=30] = "obese"

dat1$cat_sysbp = NULL
dat1$cat_sysbp[dat1$sysbp < 120] = "normal"
dat1$cat_sysbp[dat1$sysbp >=120 & dat1$sysbp < 140] = "moderate"
dat1$cat_sysbp[dat1$sysbp >=140] = "high"

## to see the first 10 obs ##
dat1[1:10, c("bmi", "cat_bmi", "sysbp", "cat_sysbp")]
```

```
> ## to see the first 10 obs ##
> dat1[1:10,c("bmi", "cat_bmi", "sysbp", "cat_sysbp")]
     bmi    cat_bmi sysbp cat_sysbp
1  26.97 overweight 106.0    normal
2     NA       <NA> 121.0  moderate
3  28.73 overweight 121.0  moderate
4  29.43 overweight 105.0    normal
5  28.50 overweight 108.0    normal
6  25.34 overweight 127.5  moderate
7  25.34 overweight 141.0      high
8  28.58 overweight 150.0      high
9  30.18      obese 183.0      high
10 23.10     normal 130.0  moderate
```

Statistical Inference for proportions
Chi-Square Tests

- **Test of Independence:**

```
## chisquare test ##
tab2 = table(dat1$cat_bmi, dat1$cat_sysbp)
prop.table(tab2, margin=2)
chisq.test(tab2)
```

```
> ## chisquare test ##
> tab2 = table(dat1$cat_bmi, dat1$cat_sysbp)
> prop.table(tab2, margin=2)

                    high    moderate      normal
  normal      0.329199549 0.445402951 0.589031079
  obese       0.208342728 0.113280363 0.053016453
  overweight  0.453213078 0.429738933 0.330895795
  underweight 0.009244645 0.011577753 0.027056673
> chisq.test(tab2)

        Pearson's Chi-squared test

data:  tab2
X-squared = 697.31, df = 6, p-value < 2.2e-16
```

reject H_0

Power & Sample Size Determination based on Effect Measures

```
## sample size calculation for risk difference ##
power.prop.test(p1 = 0.1, p2 = 0.4, power = .80, sig.level=0.05)
```

Risk Difference (RD):

- α=5%
- Power=80%
- R_1=0.1
- RD=0.3

```
> ## sample size calculation for risk difference ##
> power.prop.test(p1 = .1, p2 = 0.4, power = .80, sig.level=0.05)

     Two-sample comparison of proportions power calculation

              n = 31.49838
             p1 = 0.1
             p2 = 0.4
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

N = 32 per group. The total N = 64.

Power & Sample Size Determination based on Effect Measures

```
## sample size calculation for relative risk  = 0.3 ##
power.prop.test(p1 = 0.1, p2 = 0.03, power = .80, sig.level=0.05)
```

Relative Risk (RR):

- α=5%
- Power=80%
- R_1=0.1
- RR=0.3

(i.e., R_2 = 0.03 b/c R_2 = RR*R_1)

```
> ## sample size calculation for relative risk  = 0.3 ##
> power.prop.test(p1 = 0.1, p2 = 0.03, power = .80, sig.level=0.05)

     Two-sample comparison of proportions power calculation

              n = 193.5171
             p1 = 0.1
             p2 = 0.03
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

N = 194 per group. The total N = 388.

Power & Sample Size Determination based on Effect Measures

Odds Ratio (OR):

- α=5%
- Power=80%
- R_1=0.1
- OR=0.3

R_2 = x/(1+x),
where x = OR*R_1/(1-R_1)

```
## sample size calculation for odds ratio  = 0.3 ##
p1 = 0.1
OR = 0.3
odds1 = p1/(1-p1)
x = OR*odds1
p2 = x/(1+x)
p2

power.prop.test(p1 = 0.1, p2 = 0.03225806, power = .80, sig.level=0.05)
```

```
> ## sample size calculation for odds ratio  = 0.3 ##
> p1 = 0.1
> OR = 0.3
> odds1 = p1/(1-p1)
> x = OR*odds1
> p2 = x/(1+x)
> p2
[1] 0.03225806
>
> power.prop.test(p1 = 0.1, p2 = 0.03225806, power = .80, sig.level=0.05)

     Two-sample comparison of proportions power calculation

              n = 210.0695
             p1 = 0.1
             p2 = 0.03225806
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

N = 211 per group. The total N = 422.

Thank you