What to do today?

1. *Introdution and Preparation*

2. **Analysis with Binary Variables (Chp 1-2)**

- **2.1 Analysis with binary variables I (Chp 1)**
  - **2.1.1 On one binary variable (Chp1.1)**
    - **2.1.1A Bernoulli and binomial distributions**
    - **2.1.1B Inference on probability of success**
    - **2.1.1C More on confidence intervals**
  - *2.1.2 On two binary variables (Chp1.2)*
- *2.2 Analysis with binary response II (Chp 2)*

2.1.1A Bernoulli and binomial distribution

- When the binary variable Y is used to formulate a chance process with two outcomes, r.v. Y's distribution is a **Bernoulli distribution**: the probability mass function (pmf) is

$$P(Y = y) = \pi^y (1-\pi)^{1-y}$$

  for $y = 0, 1$. $\pi = P(Y = 1) = P(success)$: what is π?
- Multiple observations on Y: $Y_1, \ldots, Y_n$
  - n independent Bernoulli trials $\Longrightarrow Y_1, \ldots, Y_n$ are iid:

$$L(\pi | data) = \prod_{i=1}^{n} P(Y = y_i) = \pi^{\sum_{i=1}^n y_i} (1-\pi)^{n - \sum_{i=1}^n y_i}$$

  - The MLE of π: $\hat{\pi} = \sum_{i=1}^n Y_i \big/ n$, the sample proportion of success.
- $W = \sum_{i=1}^n Y_i = Y_1 + \ldots + Y_n$, the number of successes in the n trials: $\hat{\pi} = W \big/ n$
  - What is the distribution of W?

2.1.1A Bernoulli and binomial distribution

Binomial Distribution

- *Setting.* n independent Bernoulli trials -
  - two possible outcomes for each (success, failure);
  - $\pi = P(\mathit{success})$, $1 - \pi = P(\mathit{failure})$ in each trial;
  - trials are independent

- $W = \sum_{i=1}^{n} Y_i$, number of successes out of the n trials: r.v. W has the binomial distribution $B(n, \pi)$,

  $$P(W = w) = \binom{n}{w} \pi^w (1-\pi)^{n-w} = \frac{n!}{w!(n-w)!} \pi^w (1-\pi)^{n-w}$$

  for $w = 0, 1, \ldots, n$.

- Y's distribution, the Bernoulli distribution, is $B(1, \pi)$.

Example: Vote (Democrat, Republican)
Suppose $\pi = P(Democrat) = 0.60$.
For a random sample with size $n = 5$, let w = number of Democratic votes

$$p(w) = \frac{5!}{w!(5-w)!}(.6)^w(1-.6)^{5-w}$$

for $w = 0, 1, 2, 3, 4, 5$

```
> dbinom (x = 1, size = 5, prob = 0.6)
[1] 0.0768

> dbinom(x = 0:5, size = 5, prob = 0.6)
[1] 0.01024 0.07680 0.23040 0.34560 0.25920 0.07776
```

2.1.1A Bernoulli and binomial distribution

- **Mean and Variance** of $W \sim B(n, \pi)$
  - special case: $Y \sim B(1, \pi)$

$$E(Y) = (1)P(Y = 1) + (0)P(Y = 0) = \pi;$$

$$V(Y) = (1 - \pi)^2 P(Y = 1) + (0 - \pi)^2 P(Y = 0) = \pi(1 - \pi)$$

  - $W = Y_1 + \ldots + Y_n \sim B(n, \pi)$

$$E(W) = E(Y_1) + \ldots + E(Y_n) = n\pi$$

$$V(W) = V(Y_1) + \ldots + V(Y_n) = n\pi(1 - \pi)$$

- **Normal Approximation to $B(n, \pi)$:** Suppose $W \sim B(n, \pi)$. When n is large, the distribution of W is approximately Normal with mean $\mu = n\pi$ and variance $\sigma^2 = n\pi(1 - \pi)$.

Example for The Normal Approximation to Binomial Distribution: Sample surveys show that fewer people enjoy shopping than in the past. A survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming." Suppose that exactly 60% of all adult U.S. residents would say "Agree" if asked the same question. Let W = the number in the sample who agree. Estimate the probability that 1520 or more of the sample agree.

2.1.1B Inference on probability of success

- **Modeling.** Binary variable $Y \sim B(1, \pi)$
- **Data.** iid observations $Y_1, \ldots, Y_n$
- **Goal.** Make Inference about π
  - Testing: e.g. $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$
  - Estimation: e.g. confidence interval for π
- **Procedure.**
  - Likelihood based
  - Others methods

2.1.1B Inference on probability of success

Likelihood based procedures.

- $L(\pi|data) = \prod_{i=1}^{n} P(Y = y_i) = \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n} y_i}$.
- The MLE of π: $\hat{\pi} = \sum_{i=1}^{n} Y_i \big/ n$, the sample proportion of success.
    - $E(\hat{\pi}) = \pi$: unbiased
    - $\hat{\pi} \to \pi$ almost surely as $n \to \infty$: consistent
    - $\hat{\pi} \sim N(\pi, \pi(1-\pi)\big/n)$ as $n \to \infty$: asymptotical normality

2.1.1B Inference on probability of success

Likelihood based procedures. Testing on $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$

- Wald Test: approximately under H_0

$$Z = \frac{\hat{\pi} - \pi_0}{SE(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \sim N(0,1)$$

  Given significance level of α, reject H_0 if $Z_{obs} \notin (Z_{\alpha/2}, Z_{1-\alpha/2})$.

  - type I error rate α
  - $P_{H_0}(Z \in (-Z_{1-\alpha/2}, Z_{1-\alpha/2})) = 1 - \alpha$, equivalent to

$$P_{H_0}\left(\hat{\pi} - Z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n} < \pi_0 < \hat{\pi} + Z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}\right) = 1 - \alpha$$

2.1.1B Inference on probability of success

Likelihood based procedures. Confidence interval (CI) for π

- Wald type: approximately

$$Z = \frac{\hat{\pi} - \pi}{SE(\hat{\pi})} = \frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \sim N(0,1)$$

CI for π

$$\hat{\pi} \pm Z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

with confidence level of $1-\alpha$, because of

$$P\left(\hat{\pi} - Z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n} < \pi < \hat{\pi} + Z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}\right) = 1-\alpha$$

  - Easy to implement
  - Requires large n?
  - May give CI with negative values/values larger than 1?

2.1.1B Inference on probability of success

Likelihood based procedures. Testing on $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$

- Score test: $S(\pi_0) = \frac{\partial \log L(\pi|data)}{\partial \pi}\Big|_{\pi=\pi_0} = \frac{\hat{\pi}-\pi_0}{\pi_0(1-\pi_0)\big/n}$;
  approximately under H_0

$$Z = \frac{S(\pi_0)}{\sqrt{V(S(\pi_0))}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)\big/n}} \sim N(0,1)$$

  Given significance level of α, reject H_0 if $Z_{obs} \notin (Z_{\alpha/2}, Z_{1-\alpha/2})$.

  - type I error rate α
  - $P_{H_0}(Z \in (-Z_{1-\alpha/2}, Z_{1-\alpha/2})) = 1 - \alpha$, equivalent to

$$P_{H_0}\left(\hat{\pi} - Z_{1-\alpha/2}\sqrt{\pi_0(1-\pi_0)\big/n} < \pi_0 < \hat{\pi} + Z_{1-\alpha/2}\sqrt{\pi_0(1-\pi_0)\big/n}\right) = 1-\alpha$$

- Score type (Wilson CI): approximately

$$Z = \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$$

CI for π: with confidence level of $1-\alpha$,

$$\{\pi : -Z_{1-\alpha/2} < \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} < Z_{1-\alpha/2}\} \Leftrightarrow \tilde{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi}(1-\hat{\pi}) + Z^2_{1-\alpha/2}/(4n)}$$

with $\tilde{\pi} = (w + Z^2_{1-\alpha/2}/2)/(n + Z^2_{1-\alpha/2})$.

- with large n, an approximation to Wilson CI (Agresti-Coull CI):

$$\tilde{\pi} \pm Z_{1-\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n + Z^2_{1-\alpha/2}}}$$

- To implement?
- Requires large n?
- May give CI with negative values/values larger than 1?

2.1.1B Inference on probability of success

Example. $n = 10$, $w = 4$: 95% CI of π?

- Wald-type: $\hat{\pi} \pm Z_{1-\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$

```
> w<-4
> n<-10
> alpha<-0.05
> pi.hat<-w/n

> var.wald<-pi.hat*(1-pi.hat)/n
> lower<-pi.hat - qnorm(p = 1-alpha/2) * sqrt(var.wald)
> upper<-pi.hat + qnorm(p = 1-alpha/2) * sqrt(var.wald)
> round(data.frame(lower, upper), 4)
   lower  upper
1 0.0964 0.7036
```

2.1.1B Inference on probability of success

Example. $n = 10$, $w = 4$: 95% CI of π?

- Wilson (Score-type) CI:

$$\tilde{\pi} \pm \frac{Z_{1-\alpha/2}\sqrt{n}}{n + Z^2_{1-\alpha/2}} \sqrt{\hat{\pi}(1-\hat{\pi}) + Z^2_{1-\alpha/2}/(4n)}$$

with $\tilde{\pi} = (w + Z^2_{1-\alpha/2}/2)/(n + Z^2_{1-\alpha/2})$

```
> pi.tilde<-(w + qnorm(p = 1-alpha/2)^2 / 2) / (n + qnorm(p
   = 1-alpha/2)^2)
> pi.tilde
[1] 0.4277533

> Wilson C.I.
> round(pi.tilde + qnorm(p = c(alpha/2, 1-alpha/2)) *
     sqrt(n) / (n+qnorm(p = 1-alpha/2)^2) * sqrt(pi.hat*(1-
     pi.hat) + qnorm(1-alpha/2)^2/(4*n)), 4)
[1] 0.1682 0.6873
```

2.1.1B Inference on probability of success

Example. $n = 10$, $w = 4$: 95% CI of π?

- Agresti-Coull CI: $\tilde{\pi} \pm Z_{1-\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n+Z^2_{1-\alpha/2}}}$

```
> pi.tilde<-(w + qnorm(p = 1-alpha/2)^2 / 2) / (n + qnorm(p
   = 1-alpha/2)^2)
> pi.tilde
[1] 0.4277533

>  Agresti-Coull C.I.
> var.ac<-pi.tilde*(1-pi.tilde) / (n+qnorm(p = 1-alpha/2)^2)
> round(pi.tilde + qnorm(p = c(alpha/2, 1-alpha/2)) *
    sqrt(var.ac), 4)
[1] 0.1671 0.6884
```

2.1.1B Inference on probability of success

Alternative procedures. e.g. Exact Confidence interval (CI) for π (Clopper-Pearson CI) with confidence level $1 - \alpha$:

- By the *exact* distribution of $W \sim B(n, \pi)$, with observation w,

$$\{\pi : P(W \leq w) > \alpha/2 \text{ and } P(W \geq w) > \alpha/2\}$$

Example. $n = 10$, $w = 4$: 95% CI of π?

```
> binom.confint(x=4,n=10,conf.level=1-alpha,methods= exact)
  method x  n mean     lower     upper
  exact 4 10  0.4 0.1215523 0.7376219
```

2.1.1B Inference on probability of success

Alternative procedures. e.g. Exact Confidence interval (CI) for π (Clopper-Pearson CI) with confidence level $1-\alpha$:

- By the relationship between the cumulative binomial distribution and the beta distribution, the CI is

$$B(\alpha/2; w, n-w+1) < \pi < B(1-\alpha/2; w+1, n-w)$$

Example. $n = 10$, $w = 4$: 95% CI of π?

```
> alpha<-0.05
> round(qbeta(p=c(alpha/2,1-alpha/2),shape1=c(4,4+1),shape2=
    c(10-4+1,10-4)),4)
[1] 0.1216 0.7376
```

2.1.1C More on confidence intervals

- CI vs Hypothesis Testing: *There is a duality between them!*
  - have a try to apply LRT and LRT-based CI for proportion?
- Comparing the CIs
  - Wald-type vs Score-type (Wilson) CIs, and Agresti-Coull CI
  - likehood-based vs the exact CIs

Wald CI often has poor performance in categorical data anlaysis unless n is quite large.

Example. Estimate $\pi =$ population proportion of vegetarians For n = 20, observe w = 0.

Then 95% Wald CI is $0 \pm 1.96 * 0 = (0, 0) \Longrightarrow ???$

- Note what happens with Wald CI for if $\hat{\pi} = 0$ or 1
- Actual coverage probability much less than 0.95 if near 0 or 1.

- Recall Wald 95% CI is the set of π_0 values for which p-value $> .05$ in testing $H_0 : \pi = \pi_0$ vs $H_a : \pi \neq \pi_0$ using $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$ (denominator uses estimated SE)

Example. (cont'd) To estimate the probability of being vegetarian
y = 0, n = 20: $\hat{\pi} = 0$

Score-type (Wilson) CI:
What π_0 satisfies the following?

$$\pm 1.96 = \frac{0 - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{20}}} \quad \textit{or} \quad 1.96\sqrt{\frac{\pi_0(1-\pi_0)}{20}} = |0 - \pi_0|$$

Two solutions: $\pi_0 = 0$ and $\pi_0 = .16$

$\Longrightarrow$ the 95% score CI is $(0, .16)$, more sensible than the Wald CI $(0, 0)$

What will we study in the next class?

1. *Introduction and Preparation*

2. **Analysis with Binary Variables (Chp 1-2)**

- **2.1 Analysis with binary variables I (Chp 1)**
  - *2.1.1 On one binary variable (Chp1.1)*
  - **2.1.2 On two binary variables (Chp1.2)**
    - **2.1.2A Introduction**
    - **2.1.2B Inference with two binary variables**
    - **2.1.2C Beyond binary variables**
- *2.2 Analysis with binary response II (Chp 2)*