Math507M: Statistics for Medical and Public Health Researchers

Instructor: Nan LIN
nlin@wustl.edu

Department of Mathematics, College of Arts and Sciences
Division of Biostatistics, School of Medicine
Washington University in St. Louis

Acknowledgement: These slides are made based on course materials previously used by Jeff Gill and Han Liang Gan.
Populations vs. Samples

- **Population** is the whole quantity of cases of interest for some research problem.
- Usually there is not enough money or time to investigate the whole population (all middle-aged diabetics, etc).
- So we deal with a **Sample**, which is a selected subset of the populations.
- Samples are not unique and they have different properties, which can be important.
Sampling Language

▶ An **element** is the unit of analysis, a sampling element or **sampling unit** is one that becomes part of the sample.

▶ **Sampling Frame**: the segment of the population who are *able* to participate given the sampling design.

▶ Recall the difference between sample statistics and population (descriptive) statistics.

▶ **Stratum**: a population subgroup that shares some characteristic of interest.

▶ **Sampling Bias**: occurs when the sampling frame omits relevant members of the population.
Probability Sampling

- **Probability Sampling**: sample units selected from the sampling frame according to some *probabilistic* scheme.
- **Simple Random Sample**: (SRS) each element of the sampling frame has an equal opportunity of being included in the sample.
  - Old style: random numbers table.
  - New style: computerized generate.
- **SRS advantage**: randomizes across subgroups.
- The opposite of a random sample is a *convenience sample*. 
Simple random sampling in R

- With replacement:
  ```r
dat = rnorm(100)
sample(dat, size=10, replace=T)
## [1] -0.2111696 -0.5227371 -1.5785832 0.5107210 -0.5227371 -0.2984669
## [7] 0.5322637 -0.2139016 1.0317012 -0.5227371
```

- Without replacement:
  ```r
sample(dat, size=10, replace=F)
## [1] 0.31579585 0.74699492 0.29709136 1.03170116 -0.08606996
## [6] 2.18026341 -0.07745660 1.76922682 0.97872804 0.99009543
```
Cluster Sampling

- Suppose we do not have a convenient list of sampling frame elements.
- Define natural identifiable groupings (e.g. clusters), and sample from within each cluster.
- Groups are heterogeneous but identifiable.
- Classic example: we need to poll in a city, so draw random city blocks and sample households from within the blocks.
- Advantage: overcomes list problem, clustering design very flexible.
- Disadvantage: may produce more sampling error due to multi-stage design.
Example: The Swiss municipalities population

```r
library(sampling)
data(swissmunicipalities)
names(swissmunicipalities)

# # [1] "CT"       "REG"       "COM"       "Nom"       
# [5] "HApoly"   "Surfacesbois" "Surfacescult" "Alp"       
# [9] "Airbat"   "Airind"     "P00BMTOT"  "P00BWTOT"  
# [13] "Pop020"  "Pop2040"   "Pop4065"   "Pop65P"     
# [17] "H00PTOT" "H00P01"    "H00P02"    "H00P03"     
# [21] "H00P04"  "POPTOT"    

swissmunicipalities[1:3,1:5]

# #   CT REG  COM Nom HApoly 
# # 1 1 4 261 Zurich 8781 
# # 2 25 1 6621 Geneve 1593 
# # 3 12 3 2701 Basel 2391 
```
Example: A single-stage cluster sampling

- Use the swissmunicipalities data to draw a sample of clusters
- The variable 'REG' has 7 categories in the population. It is used as clustering variable
- Assume we choose 3 clusters by simple random sampling without replacement

```r
cl = cluster(swissmunicipalities,
             clustername = c("REG"), size = 3, method = "srswor")
# extracts the observed data
dat.sample = getdata(swissmunicipalities, cl)
table(cl$REG)
```

```text
##
## 3 4 7
## 321 171 245
```
Stratified sampling

- Researcher divides the entire population into different subgroups or strata, then randomly selects the final subjects proportionally from the different strata.
  - Stratification is the process of dividing members of the population into homogeneous subgroups before sampling.
- For example, geographical regions can be stratified into similar regions by means of some known variable such as habitat type, elevation or soil type.
- Another example might be to determine the proportions of defective products being assembled in a factory. In this case sampling may be stratified by production lines, factory, etc.
Use the `swissmunicipalities` data as population for drawing a sample of units.

The variable REG has 7 categories in the population. It is used as stratification variable

Computes the population stratum sizes

```r
table(swissmunicipalities$REG)
```

```
# 1 2 3 4 5 6 7
## 589 913 321 171 471 186 245
```

Sort the data to obtain the same order of the regions in the sample

```r
data = swissmunicipalities
data = data[order(data$REG),]
```
Suppose that the sample stratum sizes are given by 
size=c(30,20,45,15,20,11,44). That is, 30 units are 
drawn in the first stratum, 20 in the second one, etc.

Within a stratum, the sampling method is simple random 
sampling without replacement (equal probability, without 
replacement)

```r
st=strata(data,stratanames=c("REG"),
           size=c(30,20,45,15,20,11,44), method="srswor")
# extracts the observed data
getdata(data, st)[1:5,1:5]
```

```
## CT COM Nom HApoly Surfacesbois
## 24 22 5938 Yverdon-les-Bains 1127 86
## 227 23 6023 Conthey 8497 2076
## 288 22 5587 Le Mont-sur-Lausanne 981 189
## 518 22 5725 Prangins 603 83
## 525 22 5611 Savigny 1600 486
```

```r
table(st$REG)
```

```
##
##  1  2  3  4  5  6  7
## 30 20 45 15 20 11 44
```
Other Sampling Plans

- Iowa Caucus toilet flushing poll.
- **Snowball Sample**: use respondents to generate other respondents.
- **Quota Sample**: First, a population is segmented into mutually exclusive sub-groups, e.g. females and males. Next, non-probabilistically sample (e.g. convenience sampling) a pre-specified number of subjects from each sub-group.
- **Purposive Sample**: A researcher chooses specific people within the population to use for a particular study or research project.
- **Convenience Sample**: easy to select, cheap. For example, internet survey.
The Standard Error

- The standard deviation describes variation in data around the mean.
- In contrast, the standard error describes how variable a statistic is conditional on the data from which it is built.
- The simplest statistic that we know of is the data mean, so let’s see how its standard error changes with differing sample sizes.
- Consider the sample mean from a normal population. The standard error of the sample mean is $s/\sqrt{n}$, where $s$ is called the sample standard deviation.
Law of Large Numbers

▶ Suppose we sample \( X_1, X_2, \ldots, X_n \) such that they are iid from a certain population with finite mean and variance (does not have to be normal).

▶ Then as the sample size \( n \) gets big, \( \bar{X} \) converges to \( \mu \), the true population mean.

▶ Consider tossing a coin \( n \) times and the sample mean is then the proportion of heads.
Central Limit Theorem

- Suppose again that we sample $X_1, X_2, \ldots, X_n$ such that they are iid from a certain population with finite mean and variance.
- Then as $n$ gets “sufficiently big” the sampling distribution of the mean, $\bar{X}$, converges to a normal form with mean $\mu$ and standard error $\sigma/\sqrt{n}$: “asymptotically normal”.
- So it doesn’t matter what the original distribution of the data are, the distribution of the mean always has this property.
- This is almost like magic, but it’s true.
sample size: \( n = 1000 \)
Normal Approximation to the Binomial Distribution

- Sample $X_1, X_2, \ldots, X_n$ that are iid from a Bernoulli distribution. Then $Y = \sum_{i=1}^{n} X_i$ is distributed binomial$(n,p)$.

- Denote the sample mean as $\hat{p}$ (rather than $\bar{X}$). From the CLT, we know that $\hat{p}$ is asymptotically normal with mean $p$ and variance $\frac{p(1-p)}{n}$.

- Note $X = np$, so by the CLT, $X$ is asymptotically normally distributed when $n$ is large.
Standard Error for Proportions

- Suppose that we have a binomial experiment with true probability of a success $p$, and a sample estimate of $\hat{p}$, then:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}. \quad (1)$$

- Example: Acupuncture study (Melchart et al 2005) with 124 patients who had tension headaches and 0.46 responded to treatment, the standard error of the proportion estimate is given by:

$$SE(\hat{p}) = \sqrt{\frac{0.46(1 - 0.46)}{124}} = 0.0447.$$
Standard Error for Counts

- Suppose that we have Poisson experiment with true intensity parameter \( \lambda \), and a sample estimate of \( \hat{\lambda} \), then:

\[
SE(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{n}}.
\] (2)

- Example: Cadaveric heart donors study (Wight et al 2004) with \( r = 1.82 \) per day, the standard error of the count estimate is given by:

\[
SE(\hat{\lambda}) = \sqrt{\frac{1.82}{731}} = 0.0498.
\] (3)
**Standard Errors of a Difference**

- Sometimes we are interested in comparing two *independent* groups by creating a single statistic that tells us how different they are.
- We also need a standard error for this statistic to make reliability statements, *with the assumption that the underlying population standard deviation of the groups is really different*.

**Summary:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Estimate</th>
<th>SE(difference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Mean</td>
<td>$\mu_1 - \mu_2$</td>
<td>$\bar{x}_1 - \bar{x}_2$</td>
</tr>
<tr>
<td>Binomial</td>
<td>Proportion</td>
<td>$\pi_1 - \pi_2$</td>
<td>$p_1 - p_2$</td>
</tr>
<tr>
<td>Poisson</td>
<td>Rate</td>
<td>$\lambda_1 - \lambda_2$</td>
<td>$r_1 - r_2$</td>
</tr>
</tbody>
</table>
Difference in *Proportions*: Post-Natal Urinary Incontinence

- Glazener et al conducted a randomized clinical trial to compare two groups:
  - **Treatment**: pelvic floor muscle training exercises and bladder training, \( n_T = 279 \).
  - **Control**: standard care, \( n_C = 245 \).

for women with persistent incontinence 3 months post-natal.

- The outcome of interest is *no* urinary incontinence 12 months post-natal: \( p_T = \frac{111}{279} = 0.398 \), \( p_C = \frac{76}{245} = 0.310 \).
- The difference is therefore 0.088.
- The standard error of the difference is:

\[
SE(p_T - p_C) = \sqrt{\frac{0.398(1 - 0.398)}{279} + \frac{0.310(1 - 0.310)}{245}} = 0.0416
\]

- We compare 0.088 and 0.0416 to determine reliability.
Confidence Intervals for the Population Mean

- Start with a sample, $X_1, X_2, \ldots, X_n$, where $n$ is sufficiently large that we can rely on the CLT.
- The best estimate for $\mu$ is $\bar{x}$, and this will be the center of the interval.
- The margin of error is a distance to move out in both directions from the mean:
  \[ MOE = z_{\alpha/2} \times SE(\bar{x}) = z_{\alpha/2} \times \frac{s_x}{\sqrt{n}}, \]  
  where $z_{\alpha/2}$ is a standard normal quantile corresponding to our choice of $\alpha$.
- Examples: $\alpha = 0.05$ means that $z_{\alpha/2} = 1.96$, and $\alpha = 0.10$ means that $z_{\alpha/2} = 1.645$.
- The confidence interval is then created by:
  \[ CI_\alpha = [\bar{x} - MOE : \bar{x} + MOE] = [\bar{x} - z_{\alpha/2} \times SE(\bar{x}) : \bar{x} + z_{\alpha/2} \times SE(\bar{x})] \]
95% Confidence Interval for a Mean (z-score)
95% Confidence Interval: Birthweights

- Simpson (2004) reports the mean birthweights of 98 premature infants, summarized by:

\[
\bar{x} = 1.31\text{kg}, \quad SE(\bar{x}) = \frac{0.42}{\sqrt{98}} = 0.042\text{kg}. \quad (7)
\]

- The 95% CI is then:

\[
CI_{0.95} = [1.31 - 1.96(0.042), 1.31 + 1.96(0.042)] = [1.227, 1.392] \quad (8)
\]

- So we are 95% confident that this interval covers the true population value (the book does not say this quite right on page 91).
Defining the Confidence Interval for Proportions

- Assume $Y_1, \ldots, Y_n$ are outcomes from $n$ iid Bernoulli trials.
- When $np$ and $n(1-p)$ are both bigger than 5, we may use the CLT and treat $\hat{p} = \bar{Y}$ being normally distributed with mean $p$ and standard deviation $\sqrt{p(1-p)/n}$.
- We are interested in building a $100(1 - \alpha)\%$ confidence interval for the unknown $p$.
- Define
  \[
  SE(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n} \tag{9}
  \]
- We can standardize using the z-score for $\hat{p}$:
  \[
  z = \frac{\hat{p} - p}{SE(\hat{p})} \tag{10}
  \]
Confidence Interval for a Proportion: Acupuncture

- Recall the acupuncture study (Melchart et al. 2005) with 124 patients who had tension headaches and 0.46 responded to treatment, the standard error of the proportion estimate is given by:

\[ SE(\hat{p}) = \sqrt{\frac{0.46(1 - 0.46)}{124}} = 0.0447. \quad (11) \]

- We want a 95% confidence interval for \( p \), the true population proportion:

\[ CI_{95\%} = [0.46 - 1.96 \times 0.0447, 0.46 - 1.96 \times 0.0447] = [0.372, 0.548] \quad (12) \]

- Meaning that 19 times out of 20 we expect to cover the true population proportion with this calculation.
Interpreting Confidence

Which of these is the correct interpretation of a $100(1 - \alpha)\%$ confidence interval?

- An interval that has a $100(1 - \alpha)\%$ chance of containing the true value of the parameter.
- An interval that over $100(1 - \alpha)\%$ of replications contains the true value of the parameter, *on average*.

What interpretation do people really want.
Confidence Intervals

θ
Interpreting Confidence

```r
theta <- 5; pop <- rnorm(10000,mean=theta,sd=3)
for (i in 1:10) {
  samp <- sample(pop,30)
  moe <- 1.96*sd(samp)/sqrt(30)
  print(c(mean(samp) - moe, mean(samp) + moe))
}
```

```r
## [1] 4.501022 6.676083
## [1] 3.641503 6.201571
## [1] 4.252150 6.145494
## [1] 3.596170 6.195202
## [1] 4.257648 6.437482
## [1] 4.475340 7.304463
## [1] 4.544929 6.848892
## [1] 4.365675 6.621267
## [1] 4.150161 6.502332
## [1] 3.174625 5.471546
```
Confidence Interval for a Rate: Cadaveric Heart Donors

The cadaveric heart donors study (Wight et al 2004) had a rate of \( r = 1.82 \) per day, and the standard error of the count estimate:

\[
SE(r) = \sqrt{1.82/731} = 0.0498. \tag{13}
\]

A 95% confidence interval for the population rate \( \lambda \) is given by:

\[
CI_{0.95} = [1.82 \pm 1.96(0.0498)] = [1.7223 : 1.9176], \tag{14}
\]

which is quite narrow.

A 99% confidence interval for the population rate \( \lambda \) is given by:

\[
CI_{0.99} = [1.82 \pm 2.576(0.0498)] = [1.6917 : 1.9483], \tag{15}
\]

which wider.

Note the trade-off between confidence level and confidence width.
Confidence Intervals for Differences: Physiotherapy for Lung Cancer Patients

- Recall: $\bar{x}_T = 211$, $s_T = 118$, and $\bar{x}_C = 123$, $s_C = 99$.
- We calculated:

$$\bar{d} = \bar{x}_T - \bar{x}_C = 88, \quad SE(\bar{d}) = \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}} = \sqrt{\frac{118^2}{93} + \frac{99^2}{91}} = 16.04.$$ 

- So the 95% CI is:

$$CI_{0.95} = [\bar{d} - 1.96 \times SE(\bar{d}) : \bar{d} + 1.96 \times SE(\bar{d})]$$
$$= [88 - 1.96(16.04) : 88 + 1.96(16.05)]$$
$$= [56.542 : 119.458].$$
Example: In the 2000 New Hampshire primary (Jan 30-31), the USA Today CNN Gallup poll asked likely voters of each party to identify their choice, producing:

<table>
<thead>
<tr>
<th>Likely Republican voters</th>
<th>Likely Democratic voters</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCain</td>
<td>Gore</td>
</tr>
<tr>
<td>44%</td>
<td>54%</td>
</tr>
<tr>
<td>Bush</td>
<td>Bradley</td>
</tr>
<tr>
<td>32%</td>
<td>42%</td>
</tr>
<tr>
<td>Forbes</td>
<td>Other/undecided</td>
</tr>
<tr>
<td>13%</td>
<td>4%</td>
</tr>
<tr>
<td>Keyes</td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>(N = 697)</td>
</tr>
<tr>
<td>Bauer</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
</tr>
<tr>
<td>Hatch</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Other/undecided</td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>(N = 888)</td>
<td></td>
</tr>
</tbody>
</table>
Understanding Margin of Error in the Media

- Recall that a margin of error is half of a 95% confidence interval, defined by:

\[ [\bar{\theta} - 1.96 \times \sqrt{\text{Var}(\theta)} : \bar{\theta} + 1.96 \times \sqrt{\text{Var}(\theta)}] \] (16)

- Where \( \text{Var}(\theta) \) comes from previous polls.
- Note that \( \bar{\theta} \) is the random quantity and \( \theta \) is fixed but unknown.
- 95% is a strong convention in media polling (19/20 times coverage).
Understanding Margin of Error in the Media

- Quote from a real newspaper article:
  
  In theory, in 19 cases out of 20, the results from such polls should differ by no more than plus or minus four to five percentage points from what would have been obtained by polling the entire population of voters.

  which is correct.

- Same article:

  When a poll has a margin of error of 3 percentage points, that means there’s a 95 percent certainty that the results would differ by no more than plus or minus 3 points from those obtained if the entire voting age population was questioned.

  which is not true because of the word *certainty*.

- But it gets worse:

  Let’s say George W. Bush is up by 5 points. It sounds like this lead well exceeds the 3-point margin of error. But in fact, Bush’s support could be off by three points in either direction. So could Al Gore’s. So the real range of the poll is anywhere from an 11-point Bush lead to a 1-point Gore lead.

  where the author assumes that the candidate’s fortunes are independent.
Consider a poll with three candidates: Bush, Gore, and other. The correct distributional assumption is *multinomial* with parameters: \([p_1, p_2, p_3]\), for the true proportion of people in each group.

Define \([s_1, s_2, s_3]\) as the *sample* proportions from a single poll.

We are interested in the difference \(s_1 - s_2\) for the two leading candidates.

The expected value of this difference is \(p_1 - p_2\) and the variance is

\[
\text{Var}(s_1 - s_2) = \text{Var}(s_1) + \text{Var}(s_2) - 2\text{Cov}(s_1, s_2) \approx \frac{s_1(1 - s_1) + s_2(1 - s_2) + 2s_1s_2}{n}
\]

where the standard deviation of the difference between the two candidates is the square root of this.

Multiplying this by 1.96 gives the margin of error at the 95% confidence level.
Understanding Margin of Error in the Media

For example, a poll with \( n = 1000 \) respondents reports \( s_{\text{Bush}} = 0.46 \), \( s_{\text{Gore}} = 0.41 \), and \( s_{\text{Other}} = 0.13 \).

The newspaper claims that there is a 5 point difference with a 3% margin of error, “so Bush is clearly in the lead.”

The actual variance is produced by:

\[
\text{Var}(s_{\text{Bush}} - s_{\text{Gore}}) = \frac{(0.46)(0.54) + (0.41)(0.59) + 2(0.46)(0.41)}{1000} = 0.0008675
\]

(18)

under the assumption that lost votes do not flow to the “other” candidate.

And the square root of this produces 0.02945335.

Finally, \( 1.96 \times 0.02945335 = 0.05772857 \).

Since 0.058 is slightly more than the difference of 0.05, Gore could actually be leading.