“95% of statistics is made up on the spot”

Anonymous
Confidence interval for count data
Confidence interval for count data

- Standard error of a count, $C$, is
  
  \[ SE = \sqrt{C} \]

- For example $5 \pm 2$ (after rounding up)

- How to find a confidence interval on $\mu$?
- Exact method: a bit complicated

\[ C = 5 \pm 2 \text{ (SE)} \]
Confidence interval for count data: hand waving

Find $\mu_1$ such that $k$ cuts off 2.5% at right

Find $\mu_2$ such that $k$ cuts off 2.5% at left

$C = 5 \pm 2$ (SE)

$C = 5^{+7}_{-3}$ (95% CI)
Confidence interval for count data: exact method

- Solving equations for Poisson cumulative distribution

```r
> poisson.test(5, conf.level = 0.95)

   Exact Poisson test

data:  5 time base: 1
number of events = 5, time base = 1, p-value = 0.00366
alternative hypothesis: true event rate is not equal to 1
95 percent confidence interval:
  1.623486 11.668332
sample estimates:
  event rate
  5
```
Count errors: example

[Graph with data points and error bars labeled as 95% CI and SE]
Confidence intervals for count data are not integer

- 95% CI for $C = 5$ is $[1.6, 11.8]$
- Shouldn’t the confidence interval be exactly integer?
- Confidence interval is not for the sample count!
- We expect the true mean to be within $[1.6, 11.8]$ with a certain confidence
- The mean in a Poisson process is not integer
- Confidence intervals are for the true mean and are not integer

$$\mu = 0.7$$
Confidence interval of the correlation coefficient
Confidence interval of the correlation coefficient

- Pearson’s correlation coefficient $r$ for a sample of pairs $(x_i, y_i)$
- It is a number between -1 and 1
- It is not enough to say “we find $r = 0.82$, therefore our samples are correlated”
- Confidence limits on $r$ or significance of correlation
Sampling distribution of the correlation coefficient

- *Gedankenexperiment*
- Consider a population of pairs of numbers \((x_i, y_i)\)
- The (unknown) population correlation coefficient, \(\rho = 0.7\)
- Draw lots of samples of pairs, size \(n\)
- Calculate the correlation coefficient for each sample
- Build a sampling distribution of the correlation coefficient

![Graph showing sampling distribution of correlation coefficients](image)
Sampling distribution of the correlation coefficient

- Sampling distribution of \( r \)
- Unknown in analytical form

- Let us transform it into a known distribution

- Fisher’s transformation:

\[
Z = \frac{1}{2} \ln \frac{1 + r}{1 - r}
\]

- Build a sampling distribution of \( Z \)

Desired fraction 95%  
Reject 2.5%  
Interval containing 95% of samples
Confidence interval of the correlation coefficient

Confidence interval of $r$

$Z = \frac{1}{\sqrt{2}} \ln \frac{1 + r}{1 - r}$

Desired fraction
95%

Reject 2.5%

Confidence interval of $Z$

$\mu = \hat{Z}, \quad \sigma = \frac{1}{\sqrt{n - 3}}$
Example: 95% confidence limits on $r$

- $n = 30$ and $r = 0.7$
- First, find

$$Z = \frac{1}{2} \ln \frac{1 + r}{1 - r} = 0.867$$

$$\sigma = \frac{1}{\sqrt{n - 3}} = 0.192$$

- $Z$ is normally distributed
- 95% CI corresponds to $Z \pm 1.96\sigma$:
  - $Z_L = Z - 1.96\sigma = 0.490$
  - $Z_U = Z + 1.96\sigma = 1.24$
Example: 95% confidence limits on $r$

- $n = 30$ and $r = 0.7$
- First, find

$$Z = \frac{1}{2} \ln \frac{1 + r}{1 - r} = 0.867$$

$$\sigma = \frac{1}{\sqrt{n - 3}} = 0.192$$

- $Z$ is normally distributed
- 95% CI corresponds to $Z \pm 1.96\sigma$:
  - $Z_L = Z - 1.96\sigma = 0.490$
  - $Z_U = Z + 1.96\sigma = 1.24$
- Now we find the corresponding limits on $r$

$$r = \frac{e^{2Z} - 1}{e^{2Z} + 1}$$

- $r_L = 0.454$
- $r_U = 0.847$
- Hence, with 95% confidence, $r = 0.7^{+0.15}_{-0.25}$
How to do this in R

```r
> r <- 0.7
> n <- 30
> Z <- 0.5 * log((1+r) / (1-r))
> Z
[1] 0.8673005
> sigma <- 1 / sqrt(n - 3)
> sigma
[1] 0.1924501
> Z95 <- qnorm(0.975)
> Z95
[1] 1.959964
> Z.limits <- c(Z - Z95 * sigma, Z + Z95 * sigma)
> Z.limits
[1] 0.4901053 1.2444958
> r.limits <- (exp(2*Z.limits) - 1) / (exp(2*Z.limits) + 1)
> r.limits
[1] 0.4543000 0.8467329
```
How to do this in R: the easy way

# generate random data (in reproducible way)
> set.seed(47)
> x <- 1:30
> y <- x + rnorm(30, 0, 7)
# correlation test to find CI
> cor.test(x, y)

        Pearson's product-moment correlation

data:  x and y
t = 5.1419, df = 28, p-value = 1.882e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
  0.4494788 0.8450094
sample estimates:
   cor
0.6968971
Example: 95% CI for correlation with $n = 6$ and $n = 30$

$r = 0.7$

$CI = [-0.26, 0.96]$  
$p = 0.12$

$CI = [0.45, 0.85]$  
$p = 2 \times 10^{-5}$
Confidence interval of a proportion
Confidence interval of a proportion

- Proportion:
  \[ \hat{p} = \frac{\hat{S}}{n} = \frac{\text{number of successes}}{\text{sample size}} \]

- Examples:
  - poll results
  - survival experiments
  - counting cells with a property

- Sample proportion, \( \hat{p} \), is an estimator of the (unknown) population proportion, \( p \)

\[
\hat{S} = 4 \\
n = 12 \\
\hat{p} = \frac{4}{12} = 0.33
\]
**Sampling distribution of a proportion**

- **Gedankenexperiment**
- Consider a population of mice where \( p = 13\% \) are immune to a certain disease

- Draw a random sample of size \( n \) and find the proportion of immune mice, \( \hat{p} \), in the sample
- Repeat 100,000 times and plot the distribution of \( \hat{p} \)

- What kind of distribution is it?
- Hint: every time you select a mouse, it can be either immune or not, with probability \( p \) or \( 1 - p \)

- Binomial distribution
  - immune = “success”, probability \( p \)
  - not immune = “failure”, probability \( 1 - p \)
- Good! Sampling distribution is known
Sampling distribution of a proportion: scaled binomial

**Absolute numbers**
- \( X \) – binomial random variable
- Mean and standard deviation
  \[ \mu = np \]
  \[ \sigma = \sqrt{np(1 - p)} \]

**Proportion**
- \( R = X/n \) – scaled binomial random variable
- Mean and standard deviation scaled by \( n \):
  \[ \mu_R = p \]
  \[ \sigma_R = \frac{p(1 - p)}{\sqrt{n}} \]
Sampling distribution of a proportion

- Width of the sampling distribution of a proportion

\[ \sigma_R = \sqrt{\frac{p(1 - p)}{n}} \]
Standard error of the mean

Hypothetical experiment

- 100,000 samples of 30 mice
- Build a distribution of sample means
- Width of this distribution is the true uncertainty of the mean

\[ \sigma_m = \frac{\sigma}{\sqrt{n}} = 0.9 \text{ g} \]

Real experiment

- 30 mice
- Measure body mass: 9.9, 14.9, ..., 33.8 g
- Find standard error

\[ SE = \frac{SD}{\sqrt{n}} = 0.87 \text{ g} \]

*SE is an approximation of \( \sigma_m \)*
Sampling distribution of a proportion

- Width of the sampling distribution of a proportion
  
  \[ \sigma_R = \sqrt{\frac{p(1-p)}{n}} \]

- Replace an unknown population parameter, \( p \), with the observed estimator, \( \hat{p} \)
  
  \[ SE_R = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

- Standard error of a proportion

- \( SE_R \) estimates the width of the sampling distribution

- Approximate 95% CI is \( 1.96 \times SE_R \)

- However, this doesn’t work for small \( n \), or when proportion is close to 0 or 1
Example in R using prop.test

```r
> prop.test(1, 10)

1-sample proportions test with continuity correction

data:  1 out of 10, null probability 0.5
X-squared = 4.9, df = 1, p-value = 0.02686
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
  0.005242302 0.458846016
sample estimates:
  p
0.1
```
Confidence intervals of a proportion

- Consider survival experiment
  - take 10 mice
  - infect with something nasty
  - apply treatment
  - count survival proportion over time

- We need errors of proportion!
Consider survival experiment:
- take 10 mice
- infect with something nasty
- apply treatment
- count survival proportion over time

95% CIs using Wald method

The bigger sample, the smaller error

Even when \( \hat{p} = 0 \), error allows for non-zero proportion

We have zombie mice!
Beware of small samples!

- When you count things, small samples are not very good
- Consider a small number $n = 10$

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Correlation $r = 0.73$</th>
<th>Proportion $\hat{S} = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% CI</td>
<td>[4.8, 18.4]</td>
<td>[0.19, 0.93]</td>
<td>[0.10, 0.61]</td>
</tr>
<tr>
<td>Half CI as a fraction of the value</td>
<td>68%</td>
<td>51%</td>
<td>71%</td>
</tr>
</tbody>
</table>

- When you do counts, you need $n > 30$
Bootstrapping
Bootstrapping

- Versatile technique used when
  - distribution of the estimator is complicated or unknown
  - for power calculations
- Approximate sampling distribution from one sample only
- Use random resampling with replacement

<table>
<thead>
<tr>
<th>19.4</th>
<th>18.2</th>
<th>11.5</th>
<th>17.2</th>
<th>25.7</th>
<th>19.2</th>
<th>21.5</th>
<th>16.7</th>
<th>15.6</th>
<th>27.7</th>
<th>14.3</th>
<th>16.3</th>
<th>M = 18.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.2</td>
<td>14.3</td>
<td>19.2</td>
<td>15.6</td>
<td>14.3</td>
<td>14.3</td>
<td>17.2</td>
<td>16.3</td>
<td>19.2</td>
<td>19.2</td>
<td>16.3</td>
<td>21.5</td>
<td>M = 18.8</td>
</tr>
<tr>
<td>14.3</td>
<td>17.2</td>
<td>18.2</td>
<td>18.2</td>
<td>18.2</td>
<td>11.5</td>
<td>14.3</td>
<td>18.2</td>
<td>17.2</td>
<td>19.4</td>
<td>11.5</td>
<td>16.3</td>
<td>M = 17.2</td>
</tr>
<tr>
<td>25.7</td>
<td>18.2</td>
<td>15.6</td>
<td>15.6</td>
<td>19.4</td>
<td>19.2</td>
<td>18.2</td>
<td>19.4</td>
<td>21.5</td>
<td>16.7</td>
<td>14.3</td>
<td>18.2</td>
<td>M = 16.2</td>
</tr>
<tr>
<td>19.2</td>
<td>21.5</td>
<td>16.7</td>
<td>17.2</td>
<td>21.5</td>
<td>18.2</td>
<td>21.5</td>
<td>17.2</td>
<td>21.5</td>
<td>15.6</td>
<td>21.5</td>
<td>21.5</td>
<td>M = 18.5</td>
</tr>
<tr>
<td>19.2</td>
<td>21.5</td>
<td>16.7</td>
<td>17.2</td>
<td>21.5</td>
<td>18.2</td>
<td>21.5</td>
<td>17.2</td>
<td>21.5</td>
<td>15.6</td>
<td>21.5</td>
<td>21.5</td>
<td>M = 19.4</td>
</tr>
</tbody>
</table>

- Repeat this many times (e.g. $10^5$) and collect all means
- Build the bootstrap distribution of the mean
Bootstrapping

95% CI on the population mean
• from t-distribution [15.7, 21.5]
• from bootstrapping [16.3, 21.2]

This is not a sampling distribution, it only approximates it.
Most statistical test functions in R provide with confidence intervals

<table>
<thead>
<tr>
<th>Quantity</th>
<th>R function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>t.test</td>
</tr>
<tr>
<td>Median</td>
<td>wilcox.test</td>
</tr>
<tr>
<td>Count</td>
<td>poisson.test</td>
</tr>
<tr>
<td>Correlation</td>
<td>cor.test</td>
</tr>
<tr>
<td>Proportion</td>
<td>prop.test</td>
</tr>
</tbody>
</table>
How to extract CI limits in R

```r
> prop.test(12, 87)

1-sample proportions test with continuity correction

data:  12 out of 87, null probability 0.5
X-squared = 44.184, df = 1, p-value = 2.989e-11
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
  0.07637419 0.23243382
sample estimates:
   p
  0.137931

# store test result in a variable
> p <- prop.test(12, 87)
# extract confidence intervals from object p
> ci <- p$conf.int
# ci is a vector of two elements: lower and upper CI limit
> ci[1]
[1] 0.07637419
> ci[2]
[1] 0.2324338
```
Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html