FIT2086 Lecture 4
Central Limit Theorem and Confidence Intervals

Daniel F. Schmidt

Faculty of Information Technology, Monash University

August 16, 2022

Outline

1. The Central Limit Theorem
   - The Central Limit Theorem

2. Confidence Intervals
   - Confidence Intervals for Normal Means
   - Approximate CIs for Sample Means

Revision from last week (1)

- We looked at problem of parameter estimation
- Method of maximum likelihood

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta}\{p(\mathbf{y} \,|\, \theta)\}$$

- Maximum likelihood estimators for the normal

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_{\mathrm{ML}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\mathrm{ML}})^2}$$

- Maximum likelihood estimator for Poisson

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Revision from last week (1)

- We looked at problem of parameter estimation
- Method of maximum likelihood

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta}\{p(\mathbf{y} \mid \theta)\}$$

- Maximum likelihood estimators for the normal

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_{\mathrm{ML}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\mathrm{ML}})^2}$$

- Maximum likelihood estimator for Poisson

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Revision from last week (1)

- We looked at problem of parameter estimation
- Method of maximum likelihood

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta}\{p(\mathbf{y} \mid \theta)\}$$

- Maximum likelihood estimators for the normal

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_{\mathrm{ML}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{\mathrm{ML}})^2}$$

- Maximum likelihood estimator for Poisson

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Revision from last week (2)

- Sampling distributions of estimators
- Bias and variance of an estimator

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta, \;\; \mathrm{Var}_\theta(\hat{\theta}) = \mathbb{V}\left[\hat{\theta}\right]$$

- Mean squared error of an estimator

$$\mathrm{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta})$$

- If $Y_1, \ldots, Y_n$ have $\mathbb{E}\left[Y_i\right] = \mu$ and $\mathbb{V}\left[Y_i\right] = \sigma^2$ then

$$b_\mu(\bar{Y}) = 0, \;\; \mathrm{Var}_\mu(\bar{Y}) = \frac{\sigma^2}{n}, \;\; \mathrm{MSE}_\mu(\bar{Y}) = \frac{\sigma^2}{n}$$

- An estimator $\hat{\theta}$ is consistent if

$$b_\theta(\hat{\theta}) \to 0, \;\; \mathrm{Var}_\theta(\hat{\theta}) \to 0,$$

as $n \to \infty$ for all θ.

Revision from last week (2)

- Sampling distributions of estimators
- Bias and variance of an estimator

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta, \;\; \mathrm{Var}_\theta(\hat{\theta}) = \mathbb{V}\left[\hat{\theta}\right]$$

- Mean squared error of an estimator

$$\mathrm{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta})$$

- If $Y_1, \ldots, Y_n$ have $\mathbb{E}\left[Y_i\right] = \mu$ and $\mathbb{V}\left[Y_i\right] = \sigma^2$ then

$$b_\mu(\bar{Y}) = 0, \;\; \mathrm{Var}_\mu(\bar{Y}) = \frac{\sigma^2}{n}, \;\; \mathrm{MSE}_\mu(\bar{Y}) = \frac{\sigma^2}{n}$$

- An estimator $\hat{\theta}$ is consistent if

$$b_\theta(\hat{\theta}) \to 0, \;\; \mathrm{Var}_\theta(\hat{\theta}) \to 0,$$

as $n \to \infty$ for all θ.

Revision from last week (2)

- Sampling distributions of estimators
- Bias and variance of an estimator

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta, \;\; \mathrm{Var}_\theta(\hat{\theta}) = \mathbb{V}\left[\hat{\theta}\right]$$

- Mean squared error of an estimator

$$\mathrm{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta})$$

- If $Y_1, \ldots, Y_n$ have $\mathbb{E}\left[Y_i\right] = \mu$ and $\mathbb{V}\left[Y_i\right] = \sigma^2$ then

$$b_\mu(\bar{Y}) = 0, \;\; \mathrm{Var}_\mu(\bar{Y}) = \frac{\sigma^2}{n}, \;\; \mathrm{MSE}_\mu(\bar{Y}) = \frac{\sigma^2}{n}$$

- An estimator $\hat{\theta}$ is consistent if

$$b_\theta(\hat{\theta}) \to 0, \;\; \mathrm{Var}_\theta(\hat{\theta}) \to 0,$$

as $n \to \infty$ for all θ.

Today's Relevant Figure (N/A)

David Blackwell (1919 - 2010). Born in Illinois, US. Awarded a PhD in mathematics in 1941, being only the 7th African American to do so. Contributed to probability theory, information theory and general statistics. Known for the Rao-Blackwell theorem. First African-American to be inducted into the National Academy of Sciences.

Outline

1. The Central Limit Theorem
   - The Central Limit Theorem

2. Confidence Intervals
   - Confidence Intervals for Normal Means
   - Approximate CIs for Sample Means

The Central Limit Theorem (1)

- We have been told that the normal distribution is important
- But why is it so central to statistics?
- This is because of a special result called the central limit theorem.
- This result says that many RVs take on normal distributions, at least in some limit
- What does this all mean?

The Central Limit Theorem (2) – Key Slide

- Simple statement of the Central Limit Theorem (CLT)
- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$
- Then for large n, the distribution of

$$S = Y_1 + Y_2 + \cdots + Y_n$$

is approximately normal distributed with mean $n\mu$ and variance $n\sigma^2$

The Central Limit Theorem (3)

- More formally, we say

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2).$$

as $n \to \infty$, where "$\xrightarrow{d}$" means "converges in distribution"

- In words, the CLT says that sums of many RVs with finite means and variances are approximately normally distributed

- The approximation gets better and better for increasing n

The CLT: Implications

- So what?
- This result helps explain why so many natural phenomena seem to be normally distributed
- Consider heights of adults in a homogenous population
  $\Rightarrow$ well approximated by a normally distribution
- Why is that?
- A persons height is determined by sum of many factors:
  - Genetic causes – millions of genetic variations
  - Dietary choices, behaviour factors
- Treating these factors as RVs, we see a persons height is composed of the effects of many RVs

The CLT and Binomial distribution (1)

- Another implication is that some distributions can be approximated by normal distribution in certain cases
- Recall the binomial distribution:

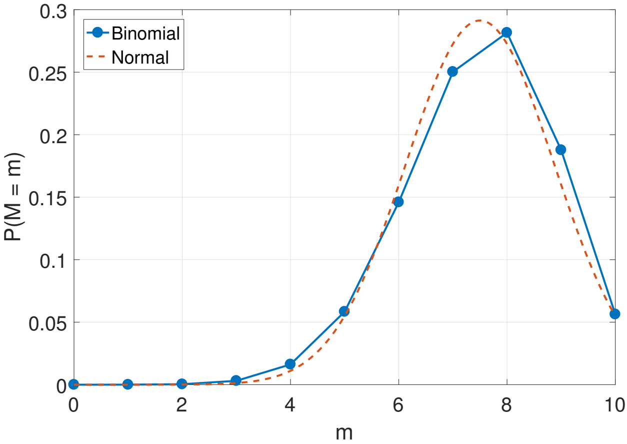
$$p(M = m \,|\, \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}$$

- This models the number of successes, M, which is defined as

$$M = \sum_{i=1}^{n} Y_i$$

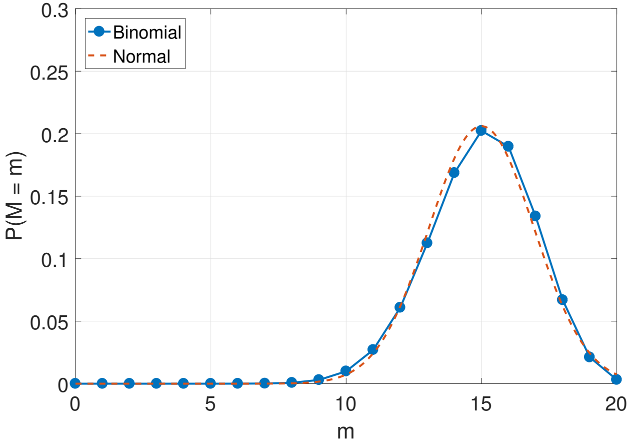
where $Y_1, \ldots, Y_n$ are RVs with $\mathbb{E}\left[Y_i\right] = \theta$, $\mathbb{V}\left[Y_i\right] = \theta(1-\theta)$
$\Rightarrow$ so by CLT, $M \sim N(n\theta, n\theta(1-\theta))$ for large n

The CLT and Binomial distribution (2)



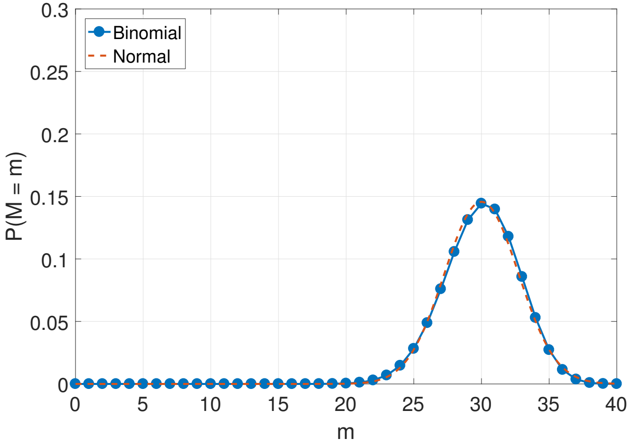
Normal $N(7.5, 1.875)$ approximation to binomial $\mathrm{Bin}(\theta = 0.75, n = 10)$ distribution.

The CLT and Binomial distribution (3)



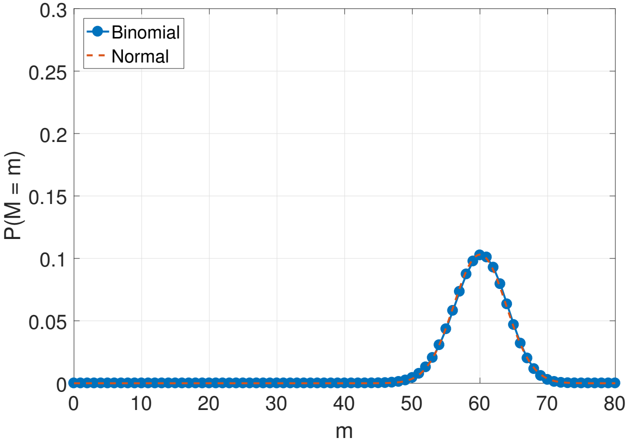
Normal $N(15, 3.75)$ approximation to binomial $\mathrm{Bin}(\theta = 0.75, n = 20)$ distribution.

The CLT and Binomial distribution (4)



Normal $N(30, 7.5)$ approximation to binomial $\mathrm{Bin}(\theta = 0.75, n = 40)$ distribution.

The CLT and Binomial distribution (5)



Normal $N(60, 15)$ approximation to binomial $\mathrm{Bin}(\theta = 0.75, n = 80)$ distribution. The two curves are now virtually the same.

The CLT and Binomial distribution (6)

- So a sum of binary RVs eventually looks normal
- Quite astonishing!
- Actually, many distributions have this property
  $\Rightarrow$ become normal as one of their parameters go to ∞
- The Poisson is another one of those that we have met ...

The CLT and Poisson distribution (1)

- Another example of this phenomena is the Poisson distribution
- For simplicity, assume λ is an integer, and consider the RV

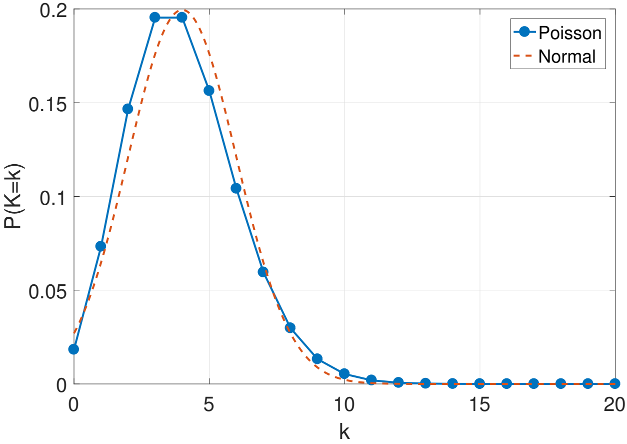
$$S \sim \mathrm{Poi}(\lambda)$$

- In Question 4.5 of Studio 2 we learned that if $X_1, \ldots, X_\lambda \sim \mathrm{Poi}(1)$ then

$$S = \sum_{i=1}^{\lambda} X_i \sim \mathrm{Poi}(\lambda)$$

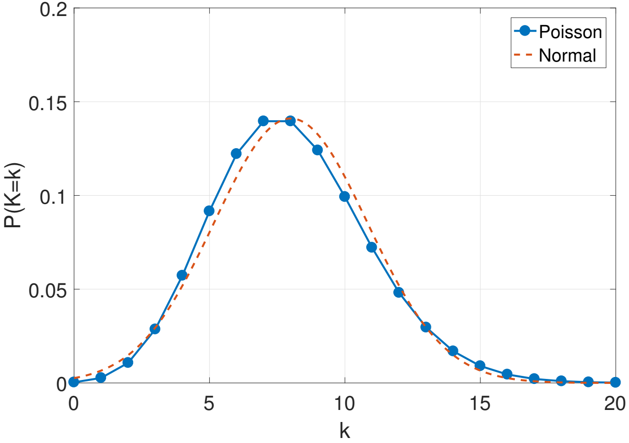
- So any $\mathrm{Poi}(\lambda)$ RV is the sum of λ $\mathrm{Poi}(1)$ RVs
- Each X_i has $\mathbb{E}\left[X_i\right] = 1$ and $\mathbb{V}\left[X_i\right] = 1$
  $\Rightarrow$ so by CLT, $S \sim N(\mu = \lambda, \sigma^2 = \lambda)$ for large λ

The CLT and Poisson distribution (2)



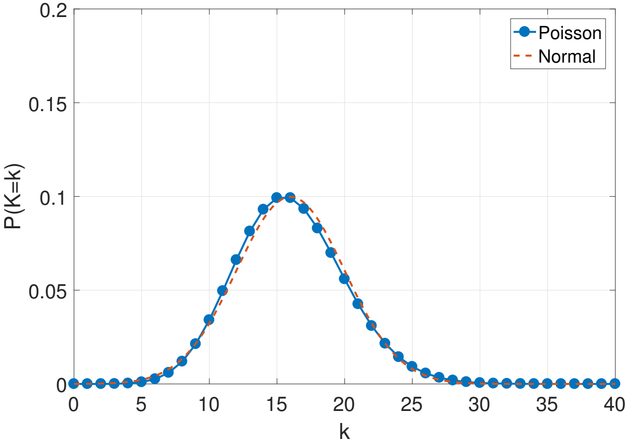
Normal $N(4, 4)$ approximation to Poisson $\mathrm{Poi}(\lambda = 4)$ distribution.

The CLT and Poisson distribution (3)



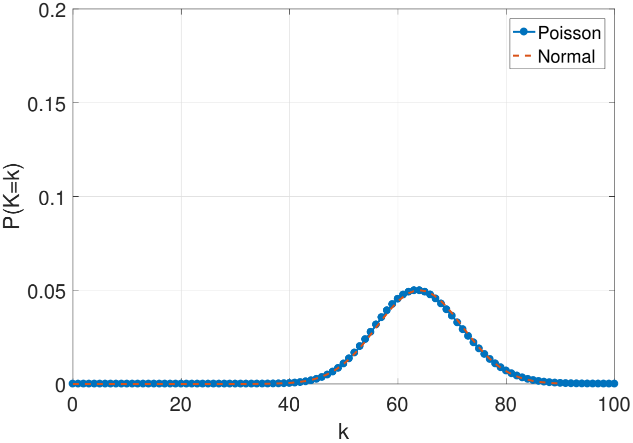
Normal $N(8, 8)$ approximation to Poisson $\mathrm{Poi}(\lambda = 8)$ distribution.

The CLT and Poisson distribution (4)



Normal $N(16, 16)$ approximation to Poisson $\mathrm{Poi}(\lambda = 16)$ distribution.

The CLT and Poisson distribution (5)



Normal $N(64, 64)$ approximation to Poisson $\mathrm{Poi}(\lambda = 64)$ distribution.

Sample means – revision (1) – Key Slide

- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs (a sample from our population)
- Assume $\mathbb{E}\left[Y_i\right] = \mu$ and $\mathbb{V}\left[Y_i\right] = \sigma^2$
- Then, the sample mean $\bar{Y}$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

  satisfies

$$\mathbb{E}\left[\bar{Y}\right] = \mu, \;\; \mathbb{V}\left[\bar{Y}\right] = \sigma^2/n$$

- In words:
  - The expected value of the sample mean is the expected value of a single datapoint from our population
  - The variance of our sample mean is the variance of a single datapoint from our population, divided by the number of datapoints in our sample

Sample means – revision (2)

- Example 1: If $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$
  - $\mathbb{E}\left[Y_i\right] = \mu$, $\mathbb{V}\left[Y_i\right] = \sigma^2$
  - So the sample mean satisfies

$$\mathbb{E}\left[\bar{Y}\right] = \mu, \;\; \mathbb{V}\left[\bar{Y}\right] = \sigma^2/n$$

- Example 2: If $Y_1, \ldots, Y_n \sim \mathrm{Poi}(\lambda)$
  - $\mathbb{E}\left[Y_i\right] = \lambda$, $\mathbb{V}\left[Y_i\right] = \lambda$
  - So the sample mean satisfies

$$\mathbb{E}\left[\bar{Y}\right] = \lambda, \;\; \mathbb{V}\left[\bar{Y}\right] = \lambda/n$$

- But what about the *distribution* of $\bar{Y}$?

CLT and Sample Means (1) – Key Slide

- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with $\mathbb{E}\left[Y_i\right] = \mu$, $\mathbb{V}\left[Y_i\right] = \sigma^2$
- From CLT we know that as $n \to \infty$

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2).$$

and $\bar{Y} = (1/n)\sum Y_i$, so using $\mathbb{V}\left[X/n\right] = \mathbb{V}\left[X\right]/n^2$ we conclude

$$\bar{Y} \xrightarrow{d} N(\mu, \sigma^2/n)$$

as $n \to \infty$

- Many estimators are an average of RVs – so very useful

CLT and Sample Means (2)

- Let $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$
  $\Rightarrow$ Then $\mathbb{E}\left[Y_i\right] = \mu$ and $\mathbb{V}\left[Y_i\right] = \sigma^2$
- From CLT we know that as $n \to \infty$

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2).$$

  and we conclude that

$$\bar{Y} \xrightarrow{d} N(\mu, \sigma^2/n)$$

  as $n \to \infty$.
- In fact, in this case the distribution of $\bar{Y}$ is exactly normal for any n

CLT and Sample Means (3)

- Another estimator of this form is

$$\hat{\lambda}_{\mathrm{ML}}(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

  which is the maximum likelihood estimator of the Poisson rate

- If $Y_1, \ldots, Y_n \sim \mathrm{Poi}(\lambda)$, then $\mathbb{E}[Y_i] = \lambda$, $\mathbb{V}[Y_i] = \lambda$, and

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\lambda, n\lambda)$$

  as $n \to \infty$, so therefore

$$\hat{\lambda}_{\mathrm{ML}} \xrightarrow{d} N(\lambda, \lambda/n)$$

- Remember, as $\hat{\lambda}_{\mathrm{ML}}$ is a sample mean its mean and variance are exactly λ and λ/n; but the *distribution* is only normal for large n

CLT and Sample Means (3)

- Another estimator of this form is

$$\hat{\lambda}_{\mathrm{ML}}(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

  which is the maximum likelihood estimator of the Poisson rate

- If $Y_1, \ldots, Y_n \sim \mathrm{Poi}(\lambda)$, then $\mathbb{E}\left[Y_i\right] = \lambda$, $\mathbb{V}\left[Y_i\right] = \lambda$, and

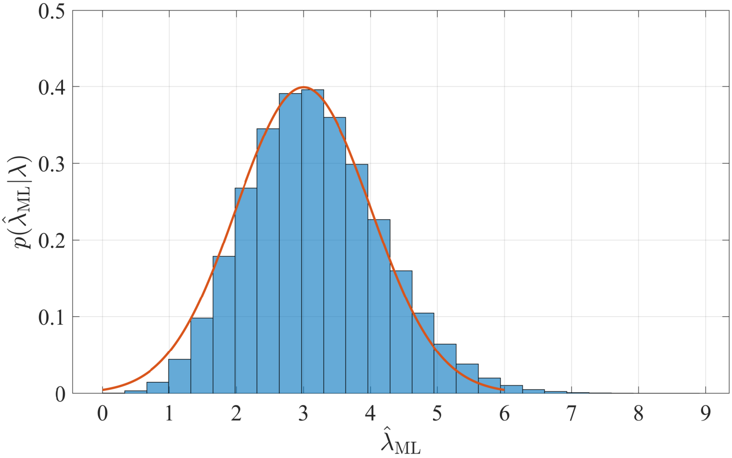
$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\lambda, n\lambda)$$

  as $n \to \infty$, so therefore

$$\hat{\lambda}_{\mathrm{ML}} \xrightarrow{d} N(\lambda, \lambda/n)$$

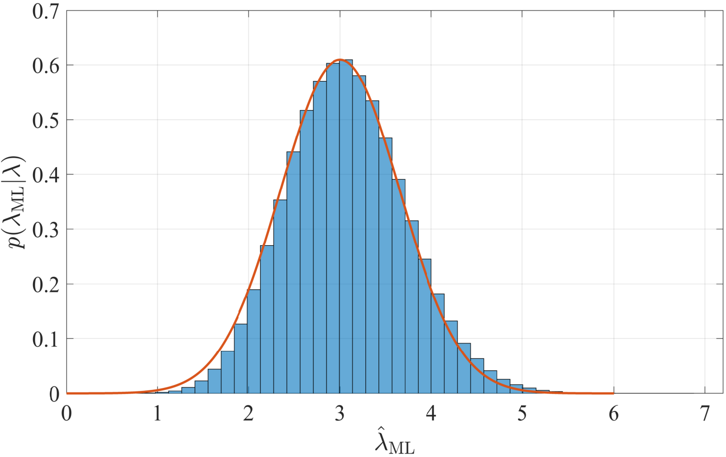
- Remember, as $\hat{\lambda}_{\mathrm{ML}}$ is a sample mean its mean and variance are exactly λ and λ/n; but the *distribution* is only normal for large n

CLT and Sample Means (4)



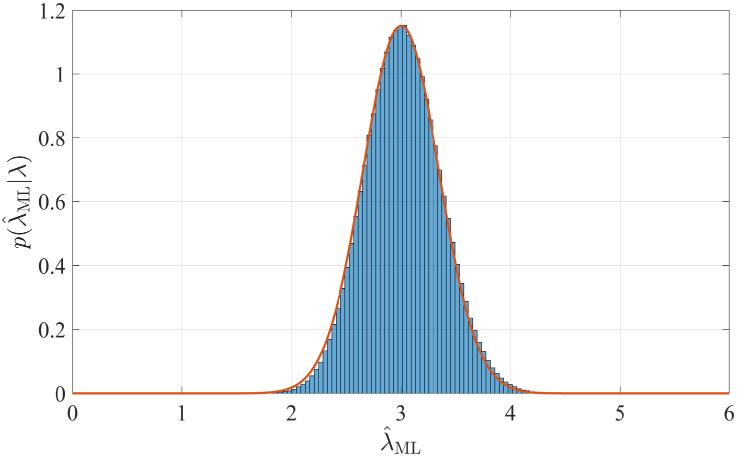
Histogram of $\hat{\lambda}_{\mathrm{ML}}$ for $n = 3$ and $\lambda = 3$; comparison with asymptotic distribution $N(3, 3/3)$.

CLT and Sample Means (5)



Histogram of $\hat{\lambda}_{\text{ML}}$ for $n = 7$ and $\lambda = 3$; comparison with asymptotic distribution $N(3, 3/7)$.

CLT and Sample Means (6)



Histogram of $\hat{\lambda}_{\mathrm{ML}}$ for $n = 25$ and $\lambda = 3$; comparison with asymptotic distribution $N(3, 3/25)$.

CLT and Sample Means (7)

- Another estimator of this form is

$$\hat{\sigma}^2_{\mathrm{ML}}(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

  which is the maximum likelihood estimator of σ^2 for a normal

- If we define $E_i = (Y_i - \bar{Y})^2$ we see it is an average of RVs

- So CLT again tells $\hat{\sigma}^2_{\mathrm{ML}}$ will be approximately normally distributed for large n

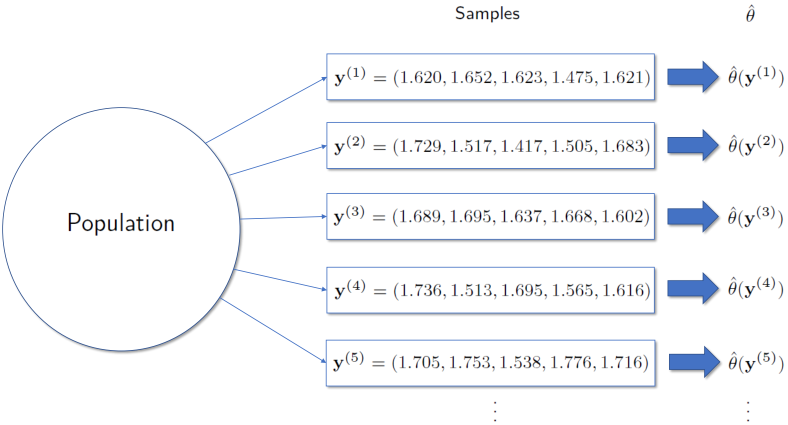
- In fact, this result holds for many estimators that don't appear on surface to be sums of RVs
  $\Rightarrow$ direct application of CLT is then difficult

Outline

1. The Central Limit Theorem
   - The Central Limit Theorem

2. Confidence Intervals
   - Confidence Intervals for Normal Means
   - Approximate CIs for Sample Means

From Population to Sample to Model – Revision



$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621)$ $\hat{\theta}(\mathbf{y}^{(1)})$

$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683)$ $\hat{\theta}(\mathbf{y}^{(2)})$

$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602)$ $\hat{\theta}(\mathbf{y}^{(3)})$

$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616)$ $\hat{\theta}(\mathbf{y}^{(4)})$

$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716)$ $\hat{\theta}(\mathbf{y}^{(5)})$

$\vdots$ $\vdots$

An (infinite) number of different random samples can be drawn from a population. Each sample would lead to a potentially different estimate $\hat{\theta}$ of a population parameter θ. The distribution of these estimates is called the sampling distribution of $\hat{\theta}$.

How to use this information? – Revision

- So now we know what the sampling distribution of an estimator (or more generally, any statistic) is.
- So what? How can we use this?
- Sampling distributions have many uses:
  - Quantifying accuracy of an estimate (confidence intervals)
  - Determining how unlikely a statistic is (hypothesis testing)
  - Comparing and evaluating quality of estimators
- Last week we examined the third use
- This week, we will look at the first

Interval Estimation (1)

- Consider a sample $\mathbf{y} = (y_1, \ldots, y_n)$
- Suppose we wish to model the population from which $\mathbf{y}$ came using a parametric distribution $p(\mathbf{y} \mid \theta)$.
- Last week we learned how to make a good guess ("estimate") a value for the parameter θ using the data
- This is called point estimation, as we estimate a single value.
- But we know our estimate is not going to be exactly correct due to randomness in our sample
- Would like to quantify how uncertain we are about the value $\Rightarrow$ this is called interval estimation.

Interval Estimation (2)

- A point estimator (like maximum likelihood) returns a single value given a sample $\mathbf{y}$, i.e., $\hat{\theta}_{\mathrm{ML}}(\mathbf{y})$
- An interval estimator returns an interval T of values, say

$$T(\mathbf{y}) = (\hat{\theta}^{-}(\mathbf{y}),\ \hat{\theta}^{+}(\mathbf{y})) \subset \mathbb{R}$$

  which says our estimate of the population parameter θ is somewhere between $\hat{\theta}^{-}(\mathbf{y})$ and $\hat{\theta}^{+}(\mathbf{y})$.

- This quantifies how uncertain we are about our estimate
  - Narrow interval $\Rightarrow$ low uncertainty
  - Wide interval $\Rightarrow$ high uncertainty
- How do we choose a good interval?

Confidence Intervals (1)

- In FIT2086, we use the method of confidence intervals.
  - From the field of frequentist statistics

- Imagine we have a recipe/procedure/algorithm for generating an interval $T(\mathbf{y})$ given an observed sample $\mathbf{y}$
  - We see a sample $\mathbf{y}$, crank the handle and $T(\mathbf{y})$ spits out some interval

- We say the interval generated by $T_\alpha(\mathbf{y})$ is a $100(1-\alpha)\%$ confidence interval, for $\alpha \in (0, 1)$, if

$$\mathbb{P}\left(\theta \in T_\alpha(\mathbf{y})\right) = 1 - \alpha,$$

- The probability is with respect to the population distribution
  - A procedure generates a $100(1-\alpha)\%$ confidence interval if for $100(1-\alpha)\%$ of possible samples from the population, the interval $T_\alpha(\mathbf{y})$ we generate contains (or "covers") the true population value

Confidence Intervals (2) – Key Slide

- In practice it is very common to consider $\alpha = 0.05$, i.e., a 95% confidence interval

- In words, imagine we have a procedure/algorithm that takes a sample $\mathbf{y}$ and returns an interval $T_{0.05}(\mathbf{y}) = (\hat{\theta}^{-}_{0.05}(\mathbf{y}),\ \hat{\theta}^{+}_{0.05}(\mathbf{y}))$

- Then, if for 95% of possible samples from the population that we could see, the interval $(\hat{\theta}^{-}_{0.05}(\mathbf{y}),\ \hat{\theta}^{+}_{0.05}(\mathbf{y}))$ generated by the procedure contains ("covers") the population value of θ, the procedure is said to generate a 95% confidence interval.

- We say: "we are 95% confident that the value of the population parameter θ lies between $\hat{\theta}^{-}_{0.05}(\mathbf{y})$ and $\hat{\theta}^{+}_{0.05}(\mathbf{y})$"

Confidence Intervals (3)

- Confidence intervals can be confusing
- They give you guarantees about a procedure/interval under repeated sampling from the population; e.g., for $\alpha = 0.05$
  - Before seeing a sample $\mathbf{y}$ from the population, we know that there is a 95% chance we will draw a sample from the population that generates a 95% confidence interval that contains ("covers") the true value of the population parameter θ
- They *do not* give you a guarantee for the particular sample you have observed
  - The population parameter θ is not a random variable – it is fixed.
  - So after observing a sample $\mathbf{y}$, the interval $(\hat{\theta}^{-}_{\alpha}(\mathbf{y}),\ \hat{\theta}^{+}_{\alpha}(\mathbf{y}))$ constructed will either contain the true value of θ, or not.

Confidence Intervals (3)

- Confidence intervals can be confusing
- They give you guarantees about a procedure/interval under repeated sampling from the population; e.g., for $\alpha = 0.05$
  - Before seeing a sample $\mathbf{y}$ from the population, we know that there is a 95% chance we will draw a sample from the population that generates a 95% confidence interval that contains ("covers") the true value of the population parameter θ
- They *do not* give you a guarantee for the particular sample you have observed
  - The population parameter θ is not a random variable – it is fixed.
  - So after observing a sample $\mathbf{y}$, the interval $(\hat{\theta}^{-}_{\alpha}(\mathbf{y}),\ \hat{\theta}^{+}_{\alpha}(\mathbf{y}))$ constructed will either contain the true value of θ, or not.

CI for Normal Mean, Known Variance (1)

- How do we generate a confidence interval?
- Let's start by constructing an CI for the mean parameter of a normal distribution
- Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ be a sample from a Gaussian population with unknown mean μ and known variance σ^2
  $\Rightarrow$ we will relax the latter assumption later on
- The maximum likelihood estimator of μ, $\hat{\mu}_{\mathrm{ML}}$, is equivalent to the sample mean

$$\hat{\mu}_{\mathrm{ML}}(\mathbf{y}) \equiv \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

CI for Normal Mean, Known Variance (2)

- Under our population assumptions, the estimate $\hat{\mu}_{\mathrm{ML}}$ is distributed as

$$\hat{\mu}_{\mathrm{ML}} \sim N(\mu, \sigma^2/n),$$

  that is, $\hat{\mu}_{\mathrm{ML}}$ exactly follows a normal distribution with mean μ and variance σ^2/n.
- We use this sampling distribution to build our 95% confidence interval

CI for Normal Mean, Known Variance (3)

- The key step is to note that

$$\frac{\hat{\mu}_{\mathrm{ML}} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

  where $\sigma/\sqrt{n}$ is the standard deviation of the estimator (square-root of the variance), and is called the standard error.

- From the above, we can then write

$$\mathbb{P}\left(-1.96 < \frac{\hat{\mu}_{\mathrm{ML}} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

  which follows from the properties of standard normal distributions (symmetry, self-similarity).

CI for Normal Mean, Known Variance (4)

- By symmetry of Gaussian distributions, multiplying through by $-\sigma/\sqrt{n}$ yields

$$\mathbb{P}\left(-1.96\frac{\sigma}{\sqrt{n}} < \mu - \hat{\mu}_{\mathrm{ML}} < \frac{\sigma}{\sqrt{n}}1.96\right) = 0.95$$

- Finally, adding $\hat{\mu}_{\mathrm{ML}}$ to all sides results in

$$\mathbb{P}\left(\hat{\mu}_{\mathrm{ML}} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu}_{\mathrm{ML}} + \frac{\sigma}{\sqrt{n}}1.96\right) = 0.95$$

which says that, for 95% of the possible samples we could draw from our population, the true population mean will be within $1.96\sigma/\sqrt{n}$ of the sample mean.

CI for Normal Mean, Known Variance (5) – Key Slide

- Assuming the population is normally distributed with (unknown) mean μ and (known) variance σ^2, these results yield the following 95% confidence interval for $\hat{\mu}_{\mathrm{ML}} \equiv \bar{Y}$,

$$\left(\hat{\mu}_{\mathrm{ML}} - 1.96\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

- More generally, a $100(1-\alpha)\%$ confidence interval is given by:

$$\left(\hat{\mu}_{\mathrm{ML}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

  where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the unit normal:
  - for $\alpha = 0.05$, $z_{0.025} = Q(p = 0.975) \approx 1, 96$;
  - for $\alpha = 0.01$, $z_{0.005} = Q(p = 0.995) \approx 2.576$;
  - for general α, use $Q(p = 1 - \alpha/2)$

  where $Q(\cdot)$ is the quantile function for the unit normal.

CI for Normal Mean, Known Variance (5) – Key Slide

- Assuming the population is normally distributed with (unknown) mean μ and (known) variance σ^2, these results yield the following 95% confidence interval for $\hat{\mu}_{\mathrm{ML}} \equiv \bar{Y}$,

$$\left(\hat{\mu}_{\mathrm{ML}} - 1.96\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

- More generally, a $100(1-\alpha)\%$ confidence interval is given by:

$$\left(\hat{\mu}_{\mathrm{ML}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

  where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the unit normal:
  - for $\alpha = 0.05$, $z_{0.025} = Q(p = 0.975) \approx 1,96$;
  - for $\alpha = 0.01$, $z_{0.005} = Q(p = 0.995) \approx 2.576$;
  - for general α, use $Q(p = 1 - \alpha/2)$

  where $Q(\cdot)$ is the quantile function for the unit normal.

CI for Normal Mean, Known Variance (6)

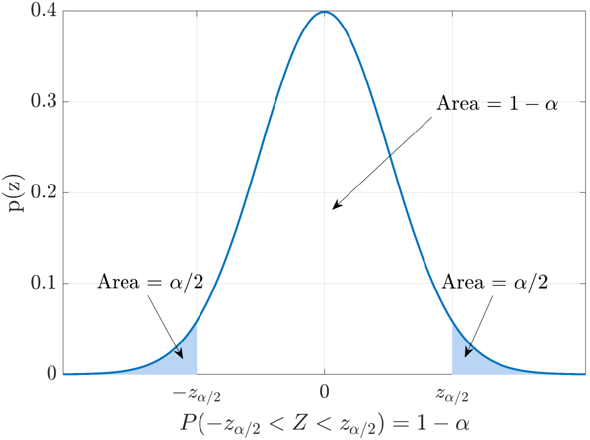
- Looking at the $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_{\mathrm{ML}}$

$$\left(\hat{\mu}_{\mathrm{ML}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

  we observe that the interval width:

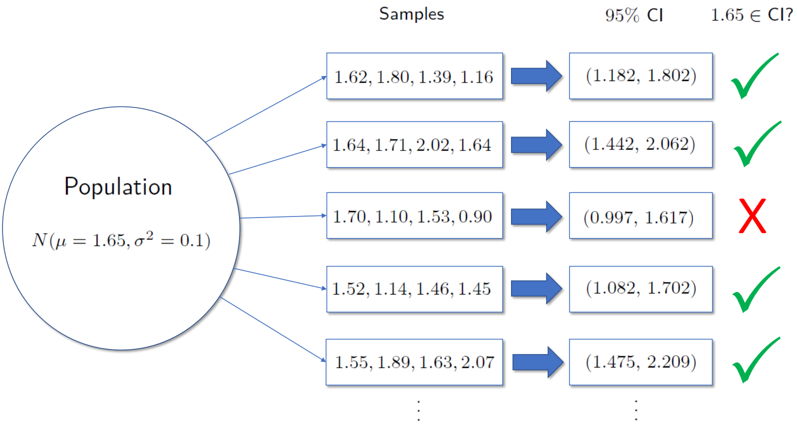
  - is proportional to the population variance σ;
  - is inversely proportional to the square-root of the sample size;
  - increases with increasing confidence level $(1-\alpha)$.

CI for Normal Mean, Known Variance (7)



Probability density of the standard normal distribution. Note that the probabilities in the tails are equal due to the symmetry of the distribution.

CI for Normal Mean, Known Variance (8)



Population

$N(\mu = 1.65, \sigma^2 = 0.1)$

Samples	95% CI	$1.65 \in$ CI?
$1.62, 1.80, 1.39, 1.16$	$(1.182,\ 1.802)$	✓
$1.64, 1.71, 2.02, 1.64$	$(1.442,\ 2.062)$	✓
$1.70, 1.10, 1.53, 0.90$	$(0.997,\ 1.617)$	✗
$1.52, 1.14, 1.46, 1.45$	$(1.082,\ 1.702)$	✓
$1.55, 1.89, 1.63, 2.07$	$(1.475,\ 2.209)$	✓
$\vdots$	$\vdots$	

Cartoon showing multiple samples drawn from a $N(\mu = 1.65, \sigma^2 = 0.1)$ population, along with the 95% confidence intervals for each sample. 5% of possible samples will result in CIs that do not include $\mu = 1.65$.

Example: Normal Mean, Known Variance (1)

- Example: We have the following samples of body mass index taken people with diabetes from the Pima ethnic group

$$\mathbf{y} = (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4)$$

- Imagine we are given a value for the population variance of 43.75 which has been estimated by another, very large study of people from the Pima group.
- Task: Estimate BMI of diabetic Pima people and construct a 95% CI
- Our best guess at the population mean BMI for Pima people with diabetes is

$$\hat{\mu}_{\rm ML} = 38.88$$

Example: Normal Mean, Known Variance (2)

- Our 95% CI is then

$$\left(38.88 - 1.96\sqrt{43.75/8},\ 38.88 + 1.96\sqrt{43.75/8}\right)$$

which is equal to

$$(34.3,\ 43.47)$$

- In words, we summarise our analysis by:

"The estimated mean BMI of people from the Pima ethnic group with diabetes (sample size $n = 8$) is $38.88\,kg/m^2$. We are 95% confident the population mean BMI for this group is between $34.3\,kg/m^2$ and $43.75\,kg/m^2$."

CI for Normal Mean, Unknown Variance (1)

- Let us make our assumptions more realistic
- $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ with *both* μ and σ^2 unknown.
- How do we construct a 95% CI for $\hat{\mu}_{\rm ML}$ in this case?
- The obvious approach would be to estimate σ^2, say using

$$\hat{\sigma}_u^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_{\rm ML})^2$$

and use this in place of the unknown variance σ^2

CI for Normal Mean, Unknown Variance (2)

- This would give a 95% CI of the form

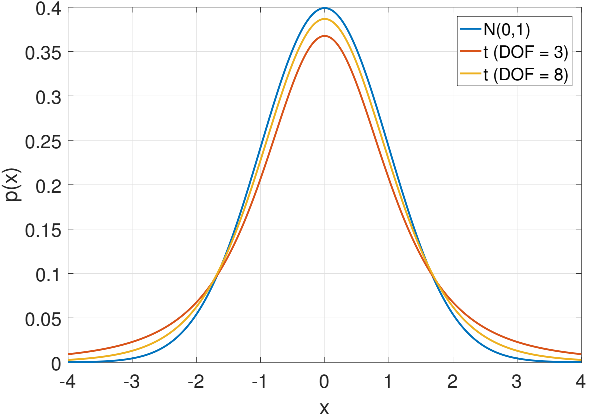
  $$\left(\hat{\mu}_{\rm ML} - 1.96\frac{\hat{\sigma}_u}{\sqrt{n}},\ \hat{\mu}_{\rm ML} + 1.96\frac{\hat{\sigma}_u}{\sqrt{n}}\right)$$

  which unfortunately, does *not* actually give 95% coverage.
- The reason is that

  $$\frac{\hat{\mu}_{\rm ML} - \mu}{\hat{\sigma}_u/\sqrt{n}}$$

  is no longer normally distributed, as the variance has been estimated from the data, rather than being known.
- It instead follows something called a Student-t distribution with $n-1$ "degrees-of-freedom"

CI for Normal Mean, Unknown Variance (3)



Plot of a standard normal $N(0,1)$ distribution and two Student-t distributions, one with degrees-of-freedom (DOF) of 3, and one with DOF of 8. Note how the t-distributions spread the probability out more and tail off to zero slower than the normal distribution.

CI for Normal Mean, Unknown Variance (4) – Key Slide

- Student-t distribution is also symmetric and self-similar, so we can instead use

$$\left(\hat{\mu}_{\mathrm{ML}} - t_{\alpha/2,n-1}\frac{\hat{\sigma}_u}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + t_{\alpha/2,n-1}\frac{\hat{\sigma}_u}{\sqrt{n}}\right)$$

  which achieves $100(1-\alpha)\%$ coverage if population is normal

- Here, $t_{\alpha/2}$ is the $100(1-\alpha/2)$-th percentile of the standard Student t-distribution with $n-1$ degrees of freedom

- To compare with normal percentiles, recall $z_{0.025} = 1.96$;
  - for $n = 3$, $t_{0.025,2} \approx 4.3$;
  - for $n = 6$, $t_{0.025,5} \approx 2.57$;
  - for $n = 11$, $t_{0.025,10} \approx 2.22$;

Example: Normal Mean, Unknown Variance (1)

- Let us revisit our Pima BMI data:

$$\mathbf{y} = (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4)$$

- This time, we do not have access to the population variance
- Our unbiased estimate of the population variance from the sample is:

$$\hat{\sigma}_u^2 = \frac{1}{7}\sum_{i=1}^{8}(y_i - 38.88)^2 \approx 51.37$$

- We also need to determine $t_{\alpha/2,n-1}$ $(\alpha = 0.05,\ n = 8)$; using R

$$\texttt{qt}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 7) \approx 2.36$$

Example: Normal Mean, Unknown Variance (2)

- This results in the 95% CI

$$\left(38.88 - 2.36\sqrt{51.37/8},\ 38.88 + 2.36\sqrt{51.37/8}\right)$$

which is equal to

$$(32.9,\ 44.86)$$

- Compare this to the “known variance” CI we obtained

$$(34.4, 43.47)$$

CI for Difference of Normal Means (1)

- Often we are interested in the difference between two samples
- Imagine we have a cohort of people in a medical trial
  - At the start of the trial, all participants' weights are measured and recorded (Sample A, population mean μ_A)
  - The participants are then administered a drug targetting weight loss
  - At the end of the trial, everyone's weight is remeasured and recorded (Sample B, population mean μ_B)
- To see if the drug had any effect, we can try to estimate the population mean difference in weights pre- and post-trial

$$\mu_A - \mu_B$$

- If no difference at population level, $\mu_A = \mu_B \Rightarrow \mu_A - \mu_B = 0$

CI for Difference of Normal Means (2)

- To estimate $\mu_A - \mu_B$, we first estimate the mean from both samples, say $\hat{\mu}_A = \bar{Y}_A$ and $\hat{\mu}_B = \bar{Y}_B$
- The estimated difference in means is then

$$\hat{\mu}_A - \hat{\mu}_B$$

- If there was no difference at a population level, we would expect on average, that $\hat{\mu}_A - \hat{\mu}_B = 0$
- But due to randomness in nature, this will never occur; so a confidence interval on $(\hat{\mu}_A - \hat{\mu}_B)$ is useful to quantify uncertainty

CI for Difference of Normal Means (3)

- Assume for the two samples A and B of size n_A and size n_B:
  - the population means μ_A and μ_B are unknown
  - the population variances σ_A^2 and σ_B^2, are known
- Then both if $\hat{\mu}_A$ and $\hat{\mu}_B$ are estimated by their respective sample means, then

$$\begin{aligned} \hat{\mu}_A &\sim N(\mu_A, \sigma_A^2/n_A) \\ \hat{\mu}_B &\sim N(\mu_B, \sigma_B^2/n_B) \end{aligned}$$

CI for Difference of Normal Means (4)

- As we assume the samples are independent, we have

$$\mathbb{V}\left[\hat{\mu}_A - \hat{\mu}_B\right] = \mathbb{V}\left[\hat{\mu}_A\right] + \mathbb{V}\left[\hat{\mu}_B\right]$$

so that the estimated difference then satisfies

$$\hat{\mu}_A - \hat{\mu}_B \sim N\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$$

- Then, we know that

$$\frac{(\hat{\mu}_A - \hat{\mu}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0, 1)$$

i.e., it follows a standard normal distribution.

CI for Difference of Normal Means (5)

- Which means the following interval

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\right)$$

  is a $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_A - \hat{\mu}_B$

- Assuming σ_A^2 and σ_B^2 known is not realistic
- If we assume they are unknown but equal, we can get exact CI on the difference (see Ross, Chapter 7.4, pp. 257-260)
  $\Rightarrow$ This is also not particularly realistic

CI for Difference of Normal Means (5)

- Which means the following interval

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \right)$$

  is a $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_A - \hat{\mu}_B$

- Assuming σ_A^2 and σ_B^2 known is not realistic
- If we assume they are unknown but equal, we can get exact CI on the difference (see Ross, Chapter 7.4, pp. 257-260)
  $\Rightarrow$ This is also not particularly realistic

CI for Difference of Normal Means (6) – Key Slide

- Instead, let us assume μ_A, μ_B, σ_A^2, σ_B^2 are all unknown
- Let $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ be unbiased estimates of the variance in sample A and B, respectively
- Then the following interval:

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}\right)$$

is an approximate $100(1-\alpha)\%$ confidence interval for $\hat{\mu}_A - \hat{\mu}_B$, with the approximation getting better for increasing n_A and n_B.

CI for Difference of Normal Means - Example (1)

- Let us return to our example involving diabete Pima people. Imagine now we have a group of non-diabetic people from the Pima group. The two samples are:

$$\begin{aligned} \mathbf{y}_N &= (34.0, 28.9, 29, 45.4, 53.2, 29.0, 36.5, 32.9) \\ \mathbf{y}_D &= (53.2, 33.6, 36.6, 42.0, 33.3, 37.8, 31.2, 43.4) \end{aligned}$$

  where $\mathbf{y}_N$ denotes non-diabetics and $\mathbf{y}_D$ denotes diabetics

- The estimates of the population mean as well as the unbiased estimates of population variance for these two groups are:

$$\begin{aligned} \hat{\mu}_N = 36.11, &\qquad \hat{\sigma}^2_N = 78.05 \\ \hat{\mu}_D = 38.88, &\qquad \hat{\sigma}^2_D = 51.37 \end{aligned}$$

CI for Difference of Normal Means - Example (2)

- The observed difference in BMI between the two groups is

$$\hat{\mu}_N - \hat{\mu}_D = 36.1 - 38.8 = -2.77\, kg/m^2$$

- The approximate 95% confidence interval is given by

$$\left(-2.77 - 1.96\sqrt{\frac{78.05}{8} + \frac{51.37}{8}},\ -2.77 + 1.96\sqrt{\frac{78.05}{8} + \frac{51.37}{8}},\right)$$

which is

$$(-10.65,\ 5.11)$$

CI for Difference of Normal Means - Example (2)

- We could summarise our results as follows:

  "The estimated difference in mean BMI between people from the Pima ethnic group without (samples size $n = 8$) and with diabetes (sample size $n = 8$) is $-2.77\,kg/m^2$. We are 95% confident the population mean difference in BMI is between $-10.65\,kg/m^2$ (BMI is lower in people without diabetes) up to $5.11\,kg/m^2$ (BMI is greater in people without diabetes). As the interval includes zero, we cannot rule out the possibility of there being no difference at a population level between people with and without diabetes."

- When looking at CI for difference, consider:
  - Interval entirely negative: suggestive of a negative difference at pop. level
  - Interval entirely positive: suggestive of a positive difference at pop. level
  - Interval contains zero: possibly no difference at pop. level

Approximate CIs for Sample Means (1)

- We have looked at CIs for the sample mean when our population is normally distributed

- But as we know, many estimators for parameters for other distributions are also the sample mean (i.e., Poisson rate, Bernoulli probability)

- In this case sampling distribution is no longer exactly normal, might even be very difficult

- We can use the central limit theorem to get approximate CIs!
  $\Rightarrow$ approximation gets better with bigger n

Approximate CIs for Sample Means (2)

- Let $Y = (Y_1, \ldots, Y_n)$ be RVs from our population
- We want to estimate some population parameter θ using Y
  - Assume only that $\mathbb{E}[Y_i] = \theta$ and $\mathbb{V}[Y_i] = v(\theta)$
- If our estimate for θ is

$$\hat{\theta}(Y) \equiv \hat{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

i.e., $\hat{\theta}$ is equivalent to the sample mean, then, from the CLT our estimate satisfies

$$\hat{\theta} \xrightarrow{d} N\left(\theta, \frac{v(\theta)}{n}\right).$$

as $n \to \infty$

Approximate CIs for Sample Means (3) – Key Slide

- This implies that as $n \to \infty$,

$$\frac{\hat{\theta} - \theta}{\sqrt{v(\theta)/n}} \xrightarrow{d} N(0,1)$$

- We don't know the true value of $v(\theta)$, but we instead use $v(\hat{\theta})$ to generate the approximate 95% confidence interval for $\hat{\theta}$

$$\left(\hat{\theta} - 1.96\sqrt{v(\hat{\theta})/n},\ \hat{\theta} + 1.96\sqrt{v(\hat{\theta})/n}\right)$$

- The quantity

$$\sqrt{v(\hat{\theta})/n}$$

is the approximate standard deviation of the estimator and is usually called the standard error of the estimate $\hat{\theta}$.

Approximate CIs for Sample Means (3) – Key Slide

- This implies that as $n \to \infty$,

$$\frac{\hat{\theta} - \theta}{\sqrt{v(\theta)/n}} \xrightarrow{d} N(0,1)$$

- We don't know the true value of $v(\theta)$, but we instead use $v(\hat{\theta})$ to generate the approximate 95% confidence interval for $\hat{\theta}$

$$\left(\hat{\theta} - 1.96\sqrt{v(\hat{\theta})/n},\ \hat{\theta} + 1.96\sqrt{v(\hat{\theta})/n}\right)$$

- The quantity

$$\sqrt{v(\hat{\theta})/n}$$

  is the approximate standard deviation of the estimator and is usually called the standard error of the estimate $\hat{\theta}$.

Example: Approximate CI for Poisson Rate Parameter

- Construct an approximate CI for the Poisson rate parameter λ
- In this case, $Y_1, \ldots, Y_n \sim \mathrm{Poi}(\lambda)$, and therefore

$$\mathbb{E}\left[Y_i\right] = \lambda, \;\; \mathbb{V}\left[Y_i\right] = v(\lambda) = \lambda$$

- The ML estimate of λ is

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$\Rightarrow$ we can use results from previous slide

- Approximate 95% CI for $\hat{\lambda}_{\mathrm{ML}}$ is then

$$\left(\hat{\lambda}_{\mathrm{ML}} - 1.96\sqrt{\hat{\lambda}_{\mathrm{ML}}/n},\; \hat{\lambda}_{\mathrm{ML}} + 1.96\sqrt{\hat{\lambda}_{\mathrm{ML}}/n}\right)$$

Example: Approximate CI for Poisson Rate Parameter

- Construct an approximate CI for the Poisson rate parameter λ
- In this case, $Y_1, \ldots, Y_n \sim \mathrm{Poi}(\lambda)$, and therefore

$$\mathbb{E}\left[Y_i\right] = \lambda, \;\; \mathbb{V}\left[Y_i\right] = v(\lambda) = \lambda$$

- The ML estimate of λ is

$$\hat{\lambda}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$\Rightarrow$ we can use results from previous slide

- Approximate 95% CI for $\hat{\lambda}_{\mathrm{ML}}$ is then

$$\left(\hat{\lambda}_{\mathrm{ML}} - 1.96\sqrt{\hat{\lambda}_{\mathrm{ML}}/n},\; \hat{\lambda}_{\mathrm{ML}} + 1.96\sqrt{\hat{\lambda}_{\mathrm{ML}}/n}\right)$$

Reading/Terms to Revise

- Reading for this week: Chapters 6 (Section 6.3) and 7 (primarily Sections 7.3, 7.4, also 7.5) of Ross.

- Terms you should know:
    - Central limit theorem;
    - Asymptotically normal;
    - Confidence interval;
    - Confidence interval of mean with known variance;
    - Confidence interval of mean with unknown variance;
    - Approximate confidence interval of difference of two means;
    - Approximate confidence interval of sample mean;

- Next week we will cover the hypothesis testing, which is related to confidence intervals.