FIT5197 Statistical Data Modelling

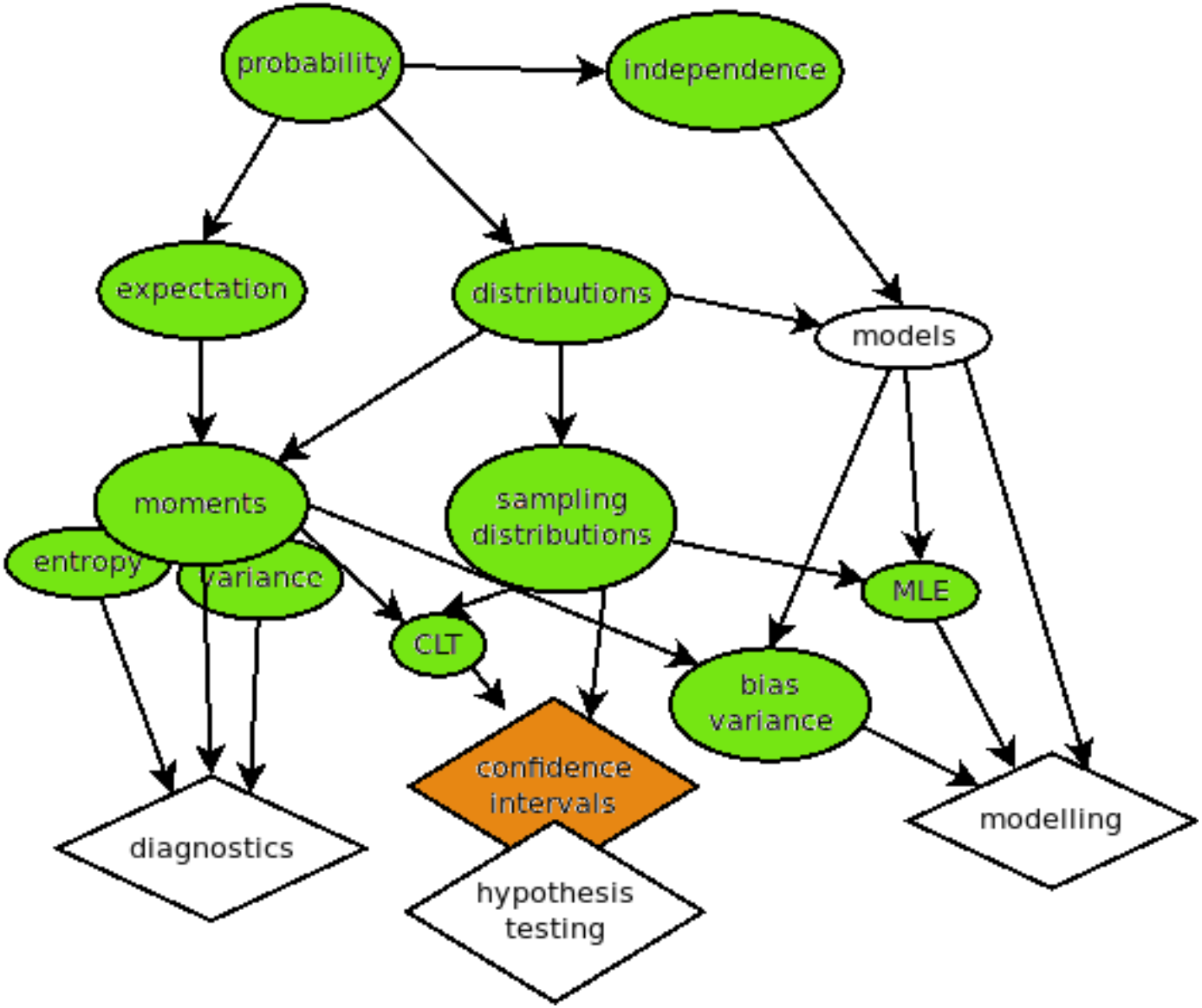
Module 3

Confidence Intervals

2020 Lecture 6

Monash University

Concept Map for This Unit



Confidence Intervals
(ePub sections 3.5
Ross 7.3-7.5)

Unit Schedule: Modules

Module	Week	Content	Ross
1.	1	introduction to modelling	1,2
2.	2	probability refresher	3
	3	random vars & expected values	4
	4	special distributions	5
3.	5	**statistical inference**	6&7
	6	**confidence intervals**	7
	7	hypothesis testing	8
4.	8	dependence & linear regression	9
	9	classification, clustering & mixtures	
5.	10	random numbers & simulation	15(bits)
	11	basic machine learning	
6.	12	modelling, validation and review	

Revision at *https://flux.qa/43FMK4*

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

Revisions(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

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

Revisions(2)

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

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

- 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 $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ then

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

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

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

as $n \to \infty$ for all θ

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

CLT

- 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?

CLT(2)

- 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 $n\sigma^2$

CLT(3)

- More formally, we say

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

  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

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?

CLT and Binomial(1)

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

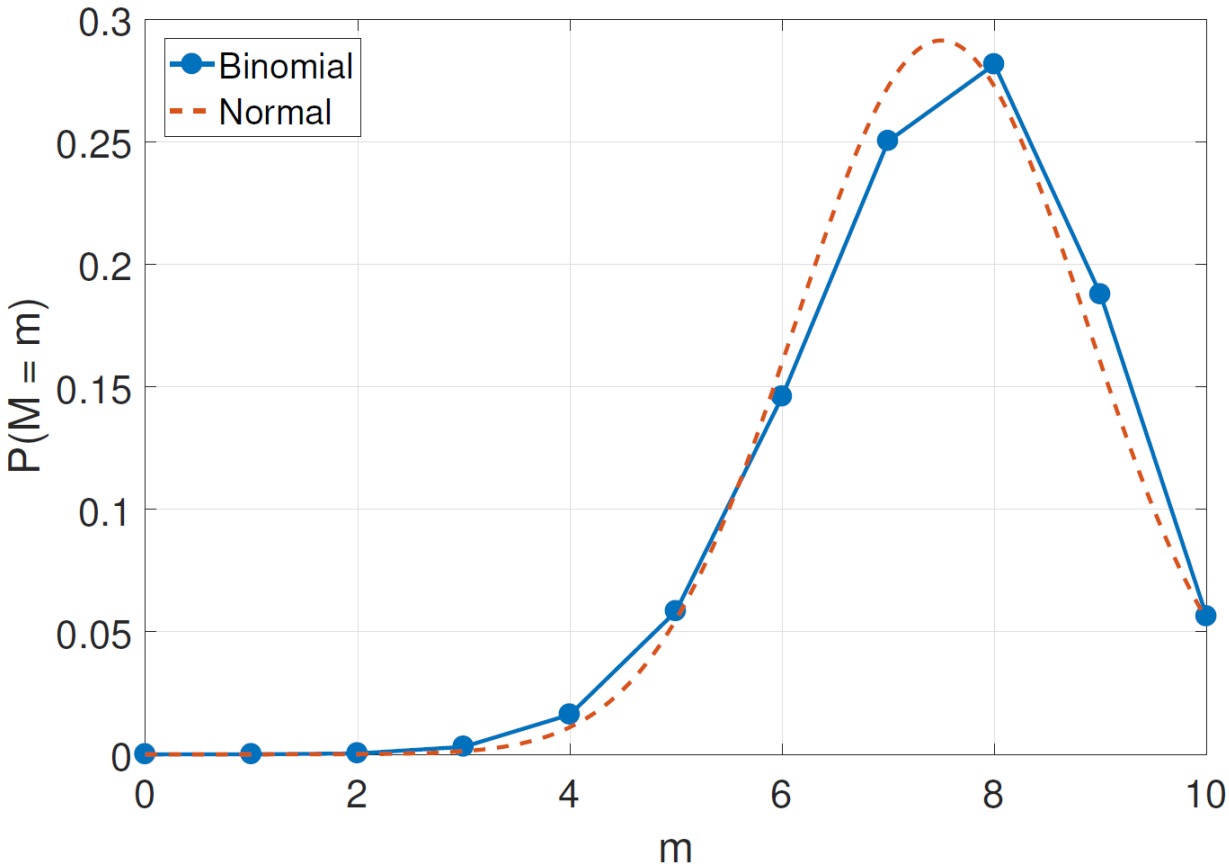
$$p(M = m|\theta) = \left(\begin{array}{c} n \\ m \end{array} \right) \theta^m (1-\theta)^{(n-m)}$$

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

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

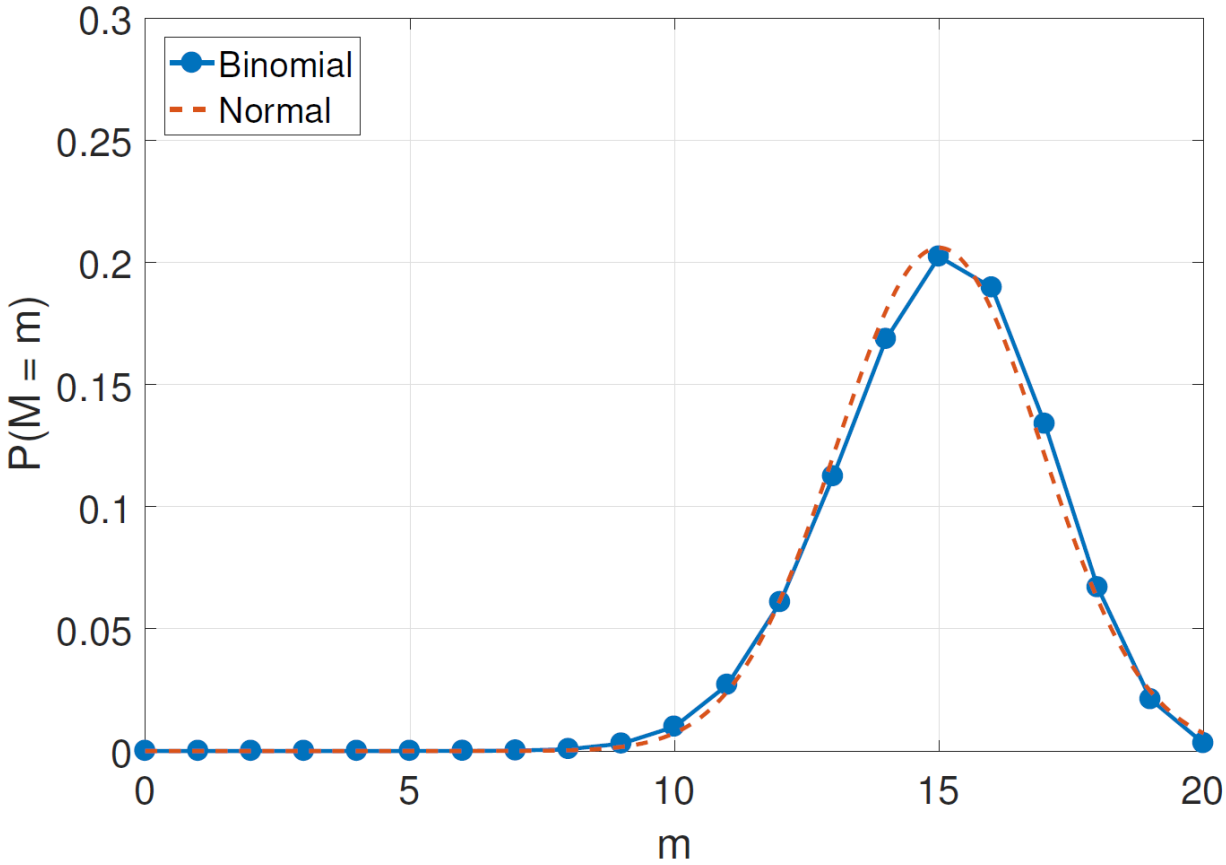
where $Y_1, \ldots, Y_n$ are RVs with $\mathbb{E}[Y_i] = \theta, \mathbb{V}[Y_i] = \theta(1-\theta)$
$\Rightarrow$ so by CLT, $M \sim N(n\theta, n\theta(1-\theta))$ for large n

CLT and Binomial(2)



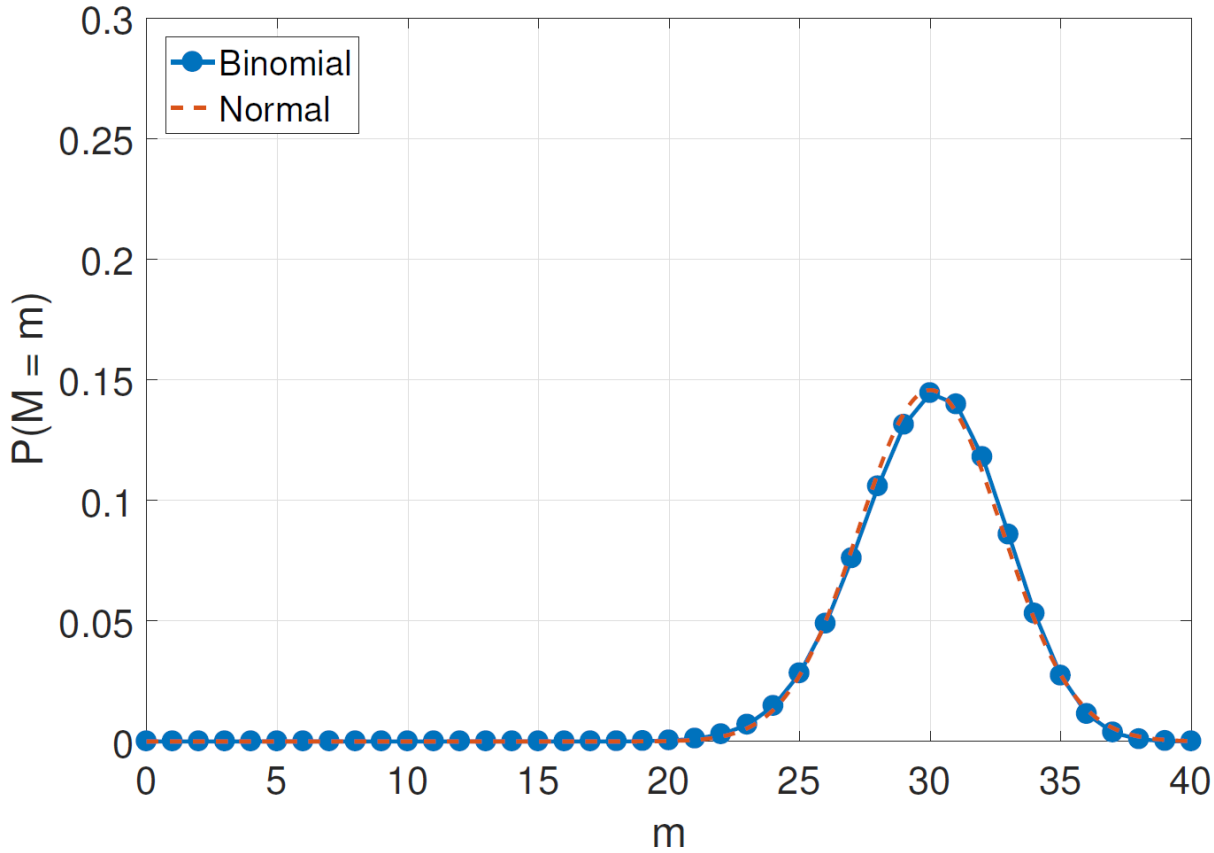
Normal $N(7.5, 1.875)$ approximation to binomial $\text{Bin}(\theta = 0.75, n = 10)$ distribution.

CLT and Binomial(3)



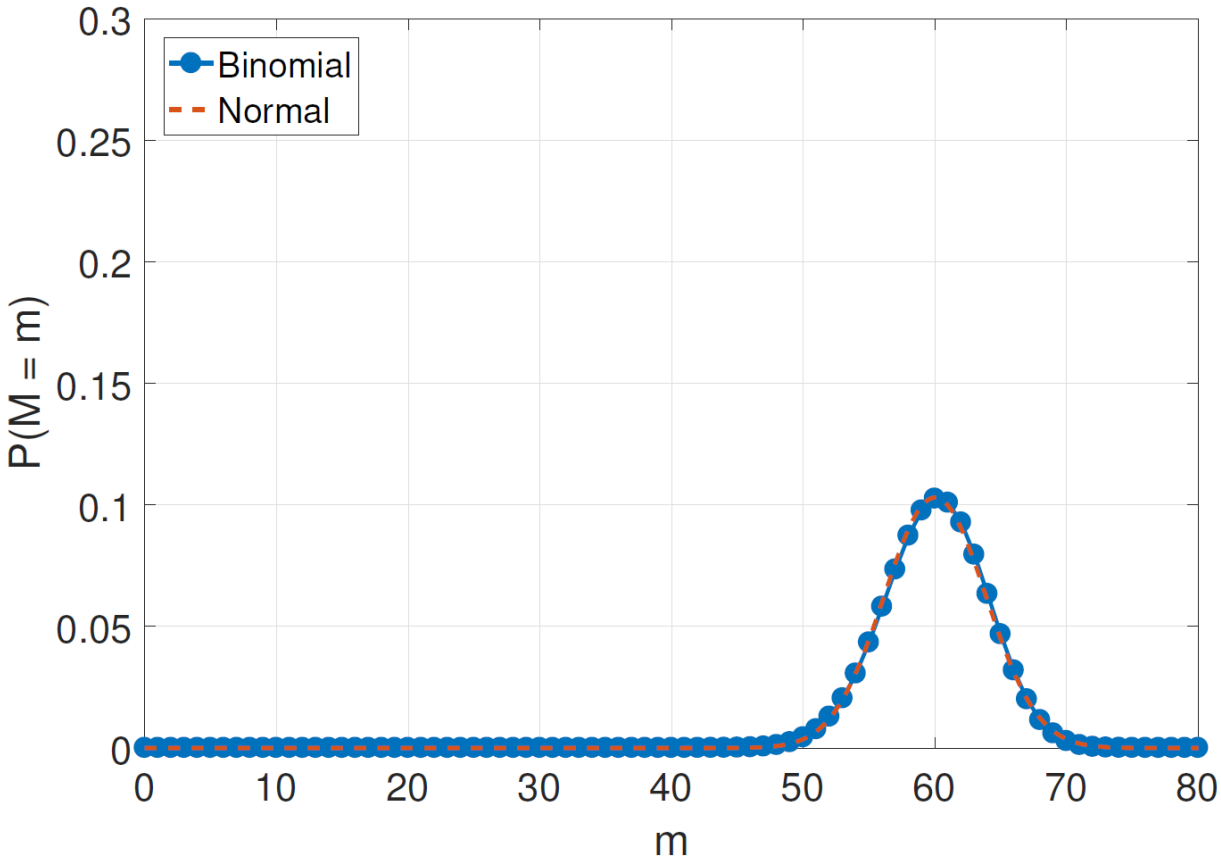
Normal $N(15, 3.75)$ approximation to binomial $\text{Bin}(\theta = 0.75, n = 20)$ distribution.

CLT and Binomial(4)



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

CLT and Binomial(5)



Normal $N(60, 15)$ approximation to binomial $\text{Bin}(\theta = 0.75, n = 80)$ distribution.

Sample means – revision (1)

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

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

satisfies

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

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 $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$

$$\mathbb{E}[Y_i] = \mu, \mathbb{V}[Y_i] = \sigma^2$$

So the sample mean satisfies

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

- Example 2: If $Y_1, \ldots, Y_n \sim \text{Poi}(\lambda)$

$$\mathbb{E}[Y_i] = \lambda, \mathbb{V}[Y_i] = \lambda$$

so the sample mean satisfies

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

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

CLT and Sample Means (1)

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

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

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

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

as $n \to \infty$

- Many estimators are an average of $\mathrm{RVs}-$ so very useful

CLT and Sample Means (2)

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

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

and we conclude that

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

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 (4)

- Another estimator of this form is

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

  which is the maximum likelihood estimator of σ^2 for a normal
- If we define $E_i = \left(Y_i - \bar{Y}\right)^2$ 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

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

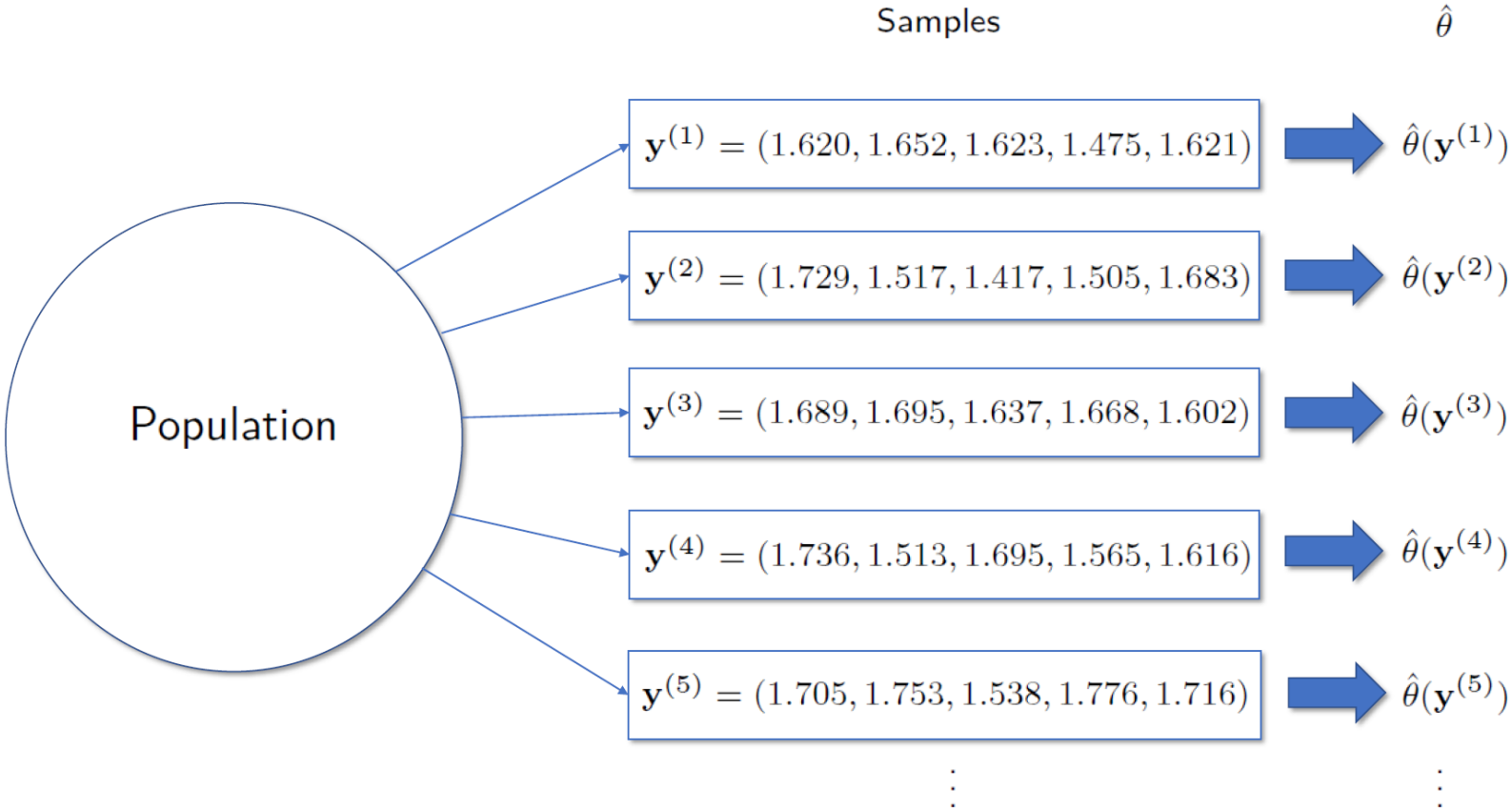
Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

Use Model-based Sampling



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 θ

How to use the information?

- 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

- 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 $p(\mathbf{y} \mid \theta)$.
- Week 5 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
  $\Longrightarrow$ this is called interval estimation.

Interval Estimation, cont.

- 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 of values, say

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

  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

- 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 $(\hat{\theta}^{-}_{0.05}(\mathbf{y}),\ \hat{\theta}^{+}_{0.05}(\mathbf{y}))$
- 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

- 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 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 y, the interval $\left(\hat{\theta}_\alpha^-(\mathbf{y}), \hat{\theta}_\alpha^+(\mathbf{y})\right)$ constructed will either contain the true value of θ, or not.

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

CI for unknown μ and known σ^2

- 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
  $\Longrightarrow$ 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 unknown μ and known σ^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 unknown μ and known σ^2

- 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

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

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

CI for unknown μ and known σ^2

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

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

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

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

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

CI for unknown μ and known σ^2

- 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 $\alpha = 0.05$, $z_{0.025} = Q(0.975) \approx 1.96$;
  - for $\alpha = 0.01$, $z_{0.005} = Q(0.995) \approx 2.576$;
  - for general α, use $Q(1-\alpha/2)$

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

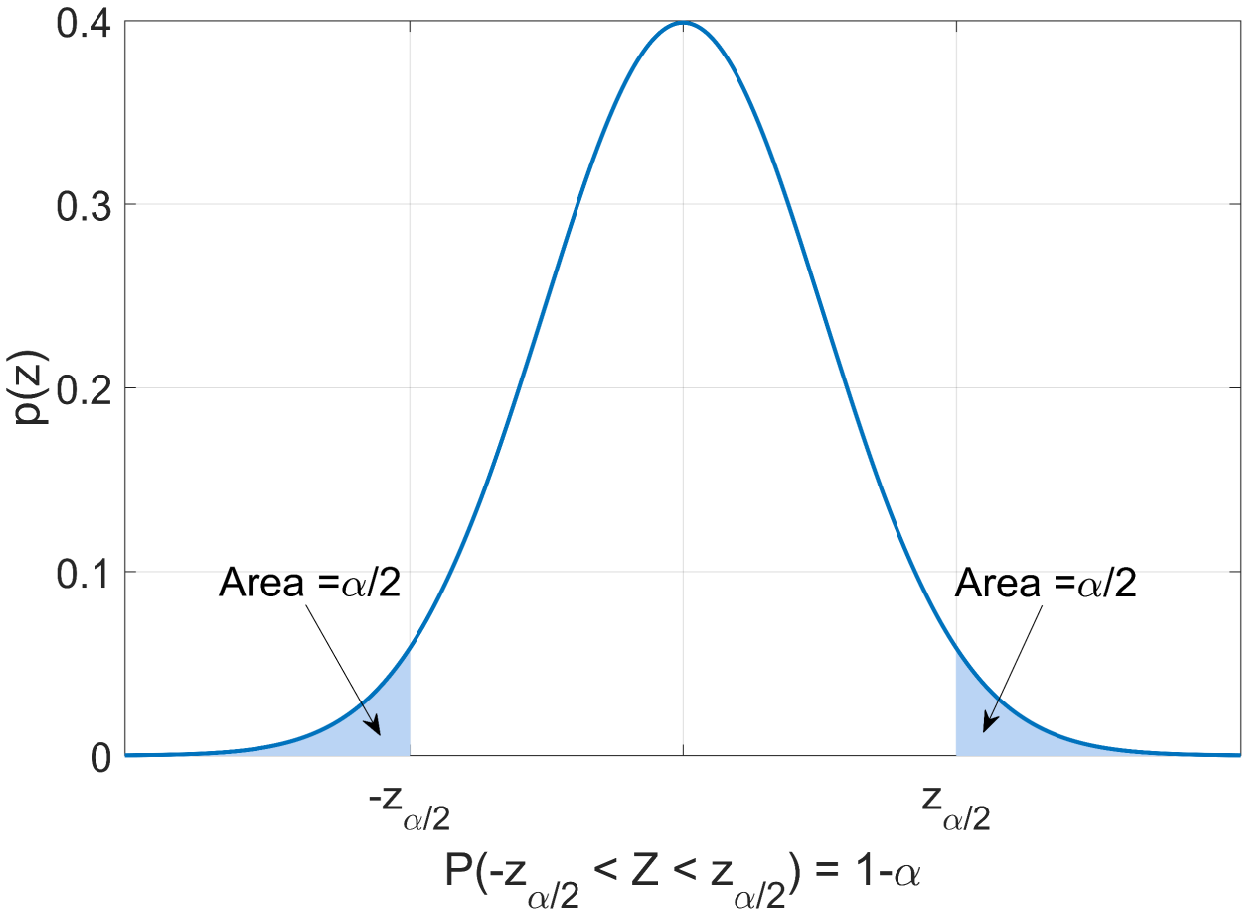
CI for unknown μ and known σ^2

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

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

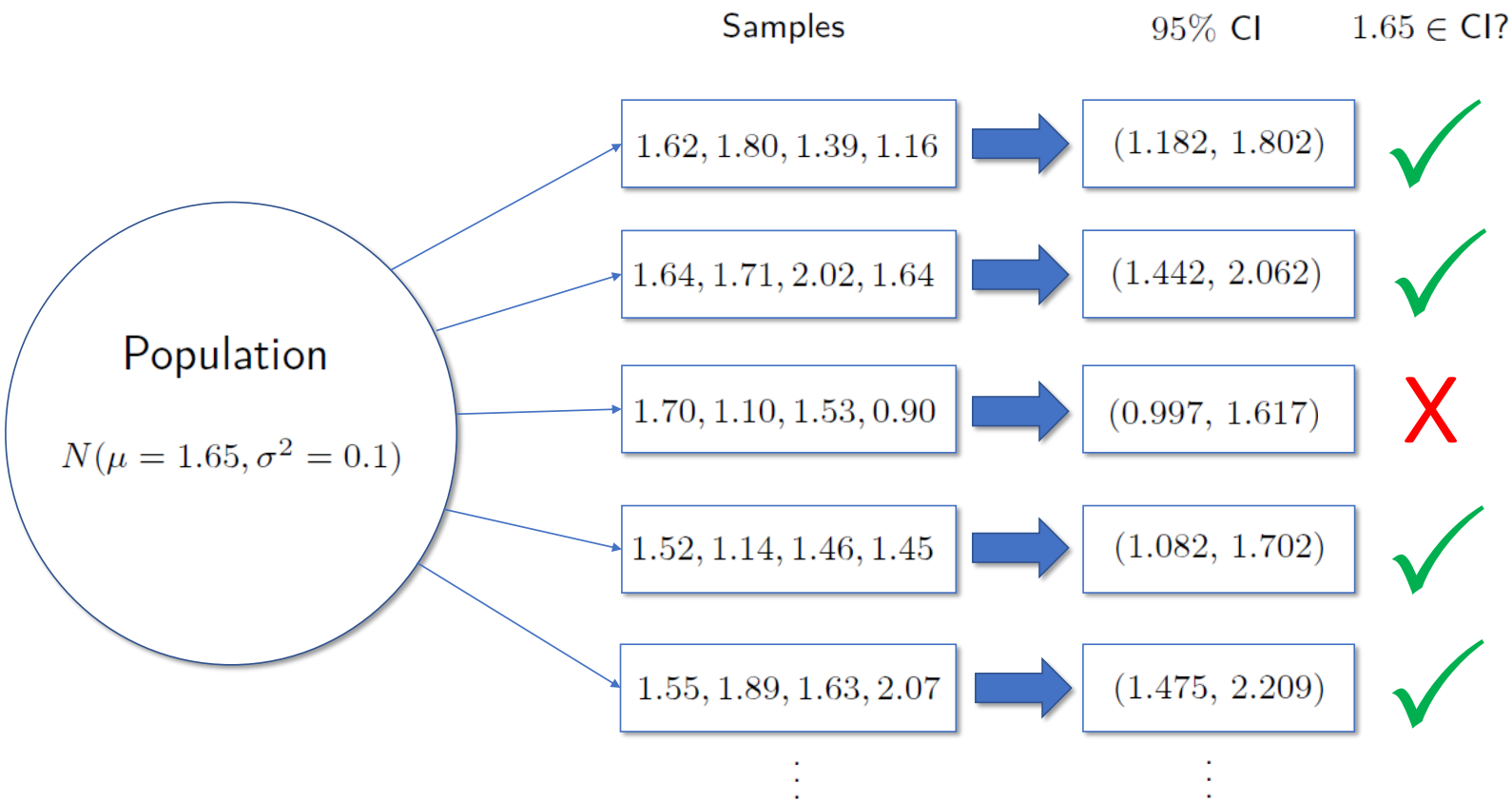
  we observe that the interval width:
  - is **proportional** to the population variance σ;
  - is **inversely proportional** to the square-root of the sample size;
  - **increases** with increasing confidence level $(1-\alpha)$.
- Do the plot showing samples being drawn with CIs

CI for unknown μ and known σ^2



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 unknown μ and known σ^2



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 $\mu = 1.65$.

Worked Example

- 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 the BMI of diabetic Pima people and construct a 95% CI
- Our best guess at the population mean BMI for Pima people with diabetes is

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

Worked Example, cont.

- 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 unknown μ and σ^2

- 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 $\hat{\mu}_{\mathrm{ML}}$ in this case?
- The obvious approach would be to estimate σ^2, say using

$$\hat{\sigma}_u^2 \equiv s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{Y})^2$$

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

CI for unknown μ and σ^2

- This would give a 95% CI of the form

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

  which unfortunately, does *not* actually give 95% coverage.

- The reason is that

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

  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 unknown μ and σ^2

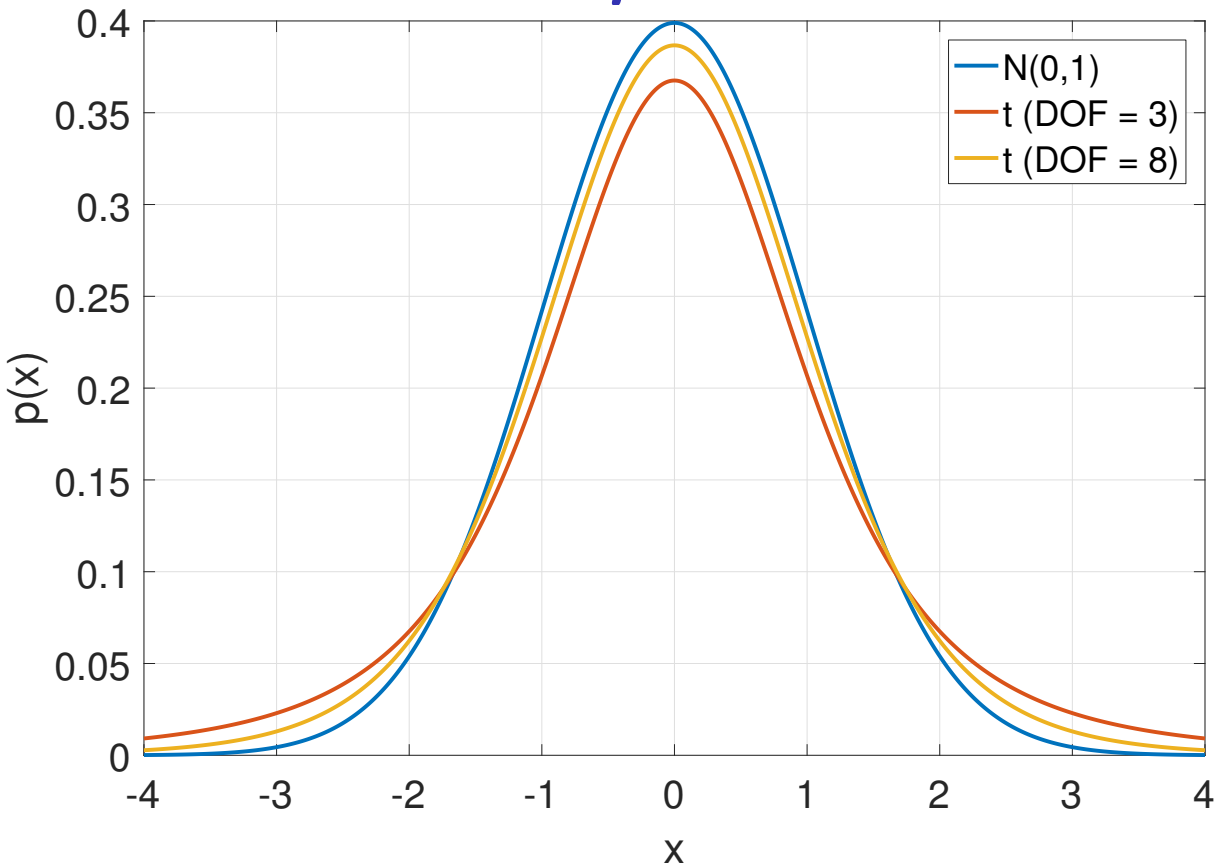


Figure: Plot of a standard normal $N(0, 1)$ distribution and two Student-t distributions, one with degrees-of-freedom (DOF) of 3, and one with DOF of 8. Note how the t-distributions spread the probability out more and tail off to zero slower than the normal distribution.

CI for unknown μ and σ^2

- 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 Gaussian

- Here, $t_{\alpha/2,n-1}$ 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 $n = 11$, $t_{0.025,10} \approx 2.22$;

  find values in *NIST critical values of Student's t distribution*

Worked Example

- 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 we find

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

Worked Example, cont.

- 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

$$(34.4, 43.47)$$

- Will the unknown variance interval always be wider?

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

CI for Difference of Normal Means

- 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 Means

- 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 Means

- 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 Means

- 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

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

follows a standard normal distribution.

CI for Difference of Means

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

CI for Difference of Means

- 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)
  $\Longrightarrow$ This is also not particularly realistic

CI for Difference of Means

- 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.

Worked Example

- 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:

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

Worked Example, cont.

- 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

$$(-10.65,\ 5.11)$$

Worked Example, cont.

- 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

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

Approximate CIs for Sample Means

- 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!
  $\Longrightarrow$ approximation gets better with bigger n

Approximate CIs for Means

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

- If our estimate for θ is

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

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

$$\hat{\theta} \xrightarrow{d} N(\theta, v(\hat{\theta})/n).$$

as $n \to \infty$

Approximate CIs for Means

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

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

Approximate CIs for Means

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

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

- 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

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

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

- The ML estimate of λ is

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

$\Longrightarrow$ we can use results from previous slide

Example: Approximate CI for Poisson Rate

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

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

- The ML estimate of λ is

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

$\Longrightarrow$ we can use results from previous slide

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

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

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

Review of Questions

Hypothesis Testing: main tool for all of empirical science

- largely recipe driven

Modelling: build a "model" for a domain problem

- to be used for prediction, "understanding," planning

Diagnostics: "debugging" a model or an algorithm

- is your model suitable for the problem?
- is the algorithm working?

Algorithm Analysis and Design: techniques and issues

- don't have to "do", but should be aware of
- building up from parts
- model "fitting", MLE, minimum cost
- bias-variance

Hypothesis Testing

Scientific Questions: have a well-defined Boolean question

(I) does drug A cause side-effect B?

(II) does Algorithm A produce better predictions than Algorithm B?

Take Measurements: take suitable measurements

(I) give some patients Drug A, give others a placebo, and measure side-effect B

(II) run both Algorithm A and Algorithm B on test data and record accuracy of predictions

Hypothesis Testing: (I) test equality of proportions in two Bernoulli populations

(II) test for difference between two Gaussian means

Hypothesis Testing

Scientific Questions: have a well-defined Boolean question

(I) does drug A cause side-effect B?
(II) does Algorithm A produce better predictions than Algorithm B?

Take Measurements: take suitable measurements

(I) give some patients Drug A, give others a placebo, and measure side-effect B
(II) run both Algorithm A and Algorithm B on test data and record accuracy of predictions

Hypothesis Testing: (I) test equality of proportions in two Bernoulli populations
(II) test for difference between two Gaussian means

generally, we need to know available recipes and choose an appropriate one for our comparison task

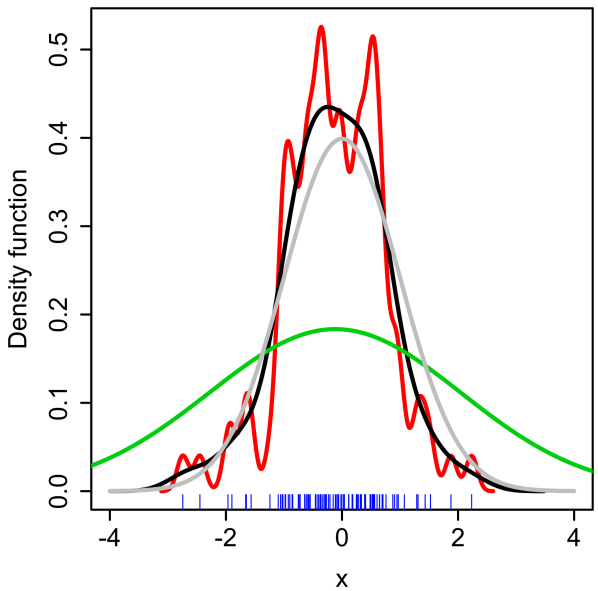
Hypothesis Testing: Example

- consider comparing a number of different prediction algorithms
- example in *Accurate parameter estimation for Bayesian network classifiers using hierarchical Dirichlet processes*
  - experiments in Table 2 on page 21
- use the *two-tail binomial sign test* "by convention"
  - why? because thats what others did

Hypothesis Testing: Example

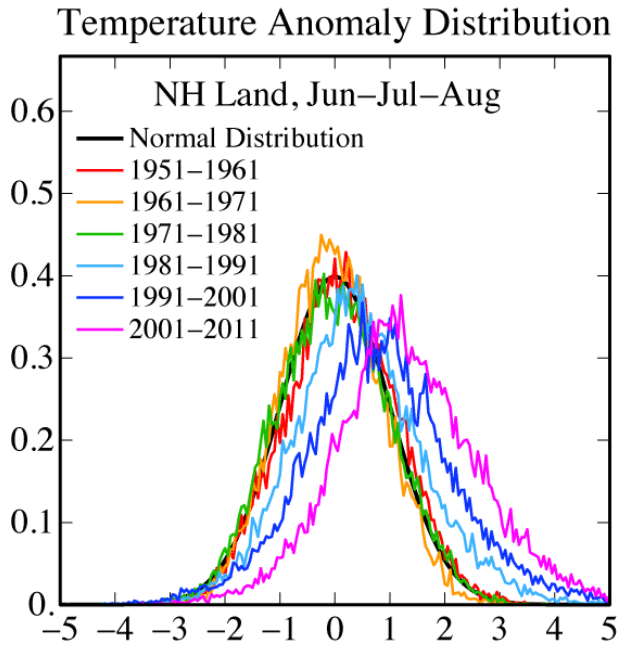
- consider comparing a number of different prediction algorithms
- example in
  Accurate parameter estimation for Bayesian network classifiers using hierarchical Dirichlet processes
  - experiments in Table 2 on page 21
- use the *two-tail binomial sign test* "by convention"
  - why? because thats what others did
  - I know little about the fine details, but I can call an R routine!
- as a data scientist, you should gradually build up a repertoire of hypothesis tests, "by example"

Modelling



by M.W.Toews [CC BY-SA 4.0], from Wikimedia Commons

what is the distribution?



NASA/GISS, Hansen, Sato and Ruedy 2012

how do temperature anomolies shift over the years?

Modelling: Readmission

	Features	Domain
Demographics	Age	in years
	Gender	Male, Female (2 categorical values)
	Marital status	Single, Married, Widowed, etc. (7 cat.)
	Ethnicity	Aboriginal, Torres Strait Islander, etc. (7 cat.)
	Birth country	154 categorical values
	Postcode	586 categorical values
Socio	Insured	binary indicator
	Health fund	Medicare Australia, Overseas, etc. (48 cat.)
Current utilization	Length of stay	in hours
	ICU stay	in minutes
	Admit ward code	193 categorical values
	Admission source	Home/Private Residence, etc. (8 cat.)
	Admission type	Admission through ED, Maternity, etc. (8 cat.)
	Admission patient classification	Public-eligible, public-ineligible, etc. (41)
	Admission specialty	Injury Cause-Poisoning, Electricity, etc. (121 cat.)
	Discharge ward code	194 categorical values
	Discharge location	194 categorical values
	Discharge method	Private residence/accommodation etc. (9 cat.)
Clinical	Primary diagnosis code	ICD-10 codes (2224 cat.)
	Primary procedure code	ICD-10 codes (1322 cat.)
	DRG code	ICD-10 codes (1008 cat.)
	Total length of stay (ever)	in hours
	Total length of stay (12 months)	in hours
	Total length of stay (6 months)	in hours

- data available for cardiac patients
- what is their probability/risk of readmission in the next 28 days after discharge?

Diagnostics: Document Modelling

- we will look at runtime diagnostics for an exploratory analysis of news articles about people in the late '80s
- answers the questions:
  - what parameters and hyperparameters were used?
  - what are the dimensions of the inputs?
  - what are the dimensions/aspects of the output
  - how well is it working?

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="2330" id="root" date="1996-08-20" xml:lang="en">
<title>USA: Tylan stock jumps; weighs sale of company.</title>
<headline>Tylan stock jumps; weighs sale of company.</headline>
<dateline>SAN DIEGO</dateline>
<text>
<p>The stock of Tylan General Inc. jumped Tuesday after the maker of
process-management equipment said it is exploring the sale of the
company and added that it has already received some inquiries from
potential buyers.</p>
<p>Tylan was up $2.50 to $12.75 in early trading on the Nasdaq market.</p>
<p>The company said it has set up a committee of directors to oversee
the sale and that Goldman, Sachs &amp;amp; Co. has been retained as its
financial adviser.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
  <code code="USA"> </code>
</codes>
<codes class="bip:industries:1.0">
  <code code="I34420"> </code>
</codes>
<codes class="bip:topics:1.0">
  <code code="C15">  </code>
  <code code="C152"> </code>
  <code code="C18">  </code>
  <code code="C181"> </code>
  <code code="CCAT"> </code>
</codes>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
<dc element="dc.date.published" value="1996-08-20"/>
<dc element="dc.source" value="Reuters"/>
<dc element="dc.creator.location" value="SAN DIEGO"/>
<dc element="dc.creator.location.country.name" value="USA"/>
<dc element="dc.source" value="Reuters"/>
</metadata>
</newsitem>
```

Figure 1: An example Reuters Corpus Volume 1 document.

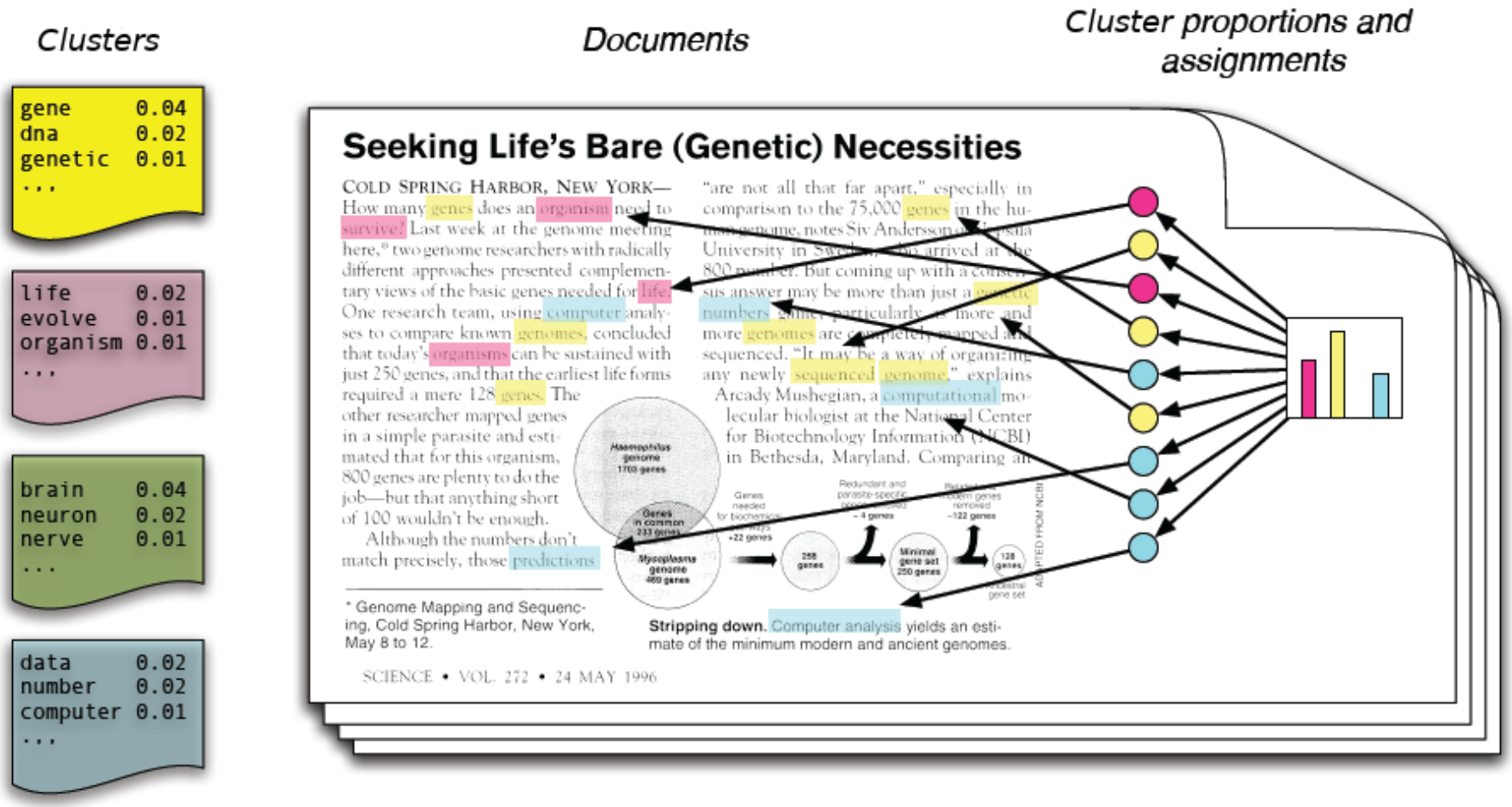
- example news article from Reuters RCV1 collection
- each document is one data record in our analysis
- want to find “word clusters” where each document can have a limited number of clusters in it

from Lewis, Yang, Rose, and Li, “RCV1: A New Benchmark Collection”, JMLR, 2004

Document Modelling

- each document is one data record in our analysis
- model a document as a small set of word clusters
- answers the questions:
  - what parameters and hyperparameters were used?
  - what are the dimensions of the inputs?
  - what word clusters were found?
  - how many word clusters are there?
  - what are their effective dimensions?

Word Clusters: Example



```
[24/09/2015:22:46:07] COMMAND-LINE: hca -q 4 -K200 -W 10000 -f ldac -t7616 -Llike,0,0 -v -v -V -Sbdk=100 -Ang -C300 pn PPE
Version 0.61, threads, Normalised Gamma sampler for topics, H.Pitman-Yor sampler for words
Setting seed = 1445687167
Read from ldac file: D=8616, W=14768, N=1665733
Sampling pars: aw(11), bw(3), aw0(11), bw0(3), bdk(3), NGalpha(4), NGbeta(4),
Sampling in batches of 40:  bdk, NGbeta, NGalpha,
mem   = 35.5 (MByte)
seed  = 1445687167
N     = 1586697
W     = 10000
D     = 8616
TRAIN   = 7616
TEST    = 0
T     = 200
ITER  = 300
PYbeta  = 2
aw     = 0.500000
bw     = 100.000000
aw0     = 0.500000
bw0     = 10.000000
PYalpha  = 0
[24/09/2015:22:46:08] cycles: cycle 0
log_2(perp)=15.5264,9.4047
Pars:  aw=0.500000, bw=460.793880, aw0=0.500000, bw0=47.179683, bdk=100.000000, NGalpha=0.045701, NGbeta=0.005000
conc. = 0.000000, empty = 0, exp.ent = 199.376429

....

Topic 0 p=5.64% ew=219.1 ed=3425.6 pmi=0.525
        words=think,n't,really,else,things,feel,thing,'ll,mind,'ve
Topic 1 p=4.10% ew=116.7 ed=431.6 pmi=0.337
        words=bypass,debakey,yastrzhembsky,yeltsin,akchurin,kremlin,mironov,65-year-old,sergei,quintuple
Topic 2 p=3.78% ew=584.7 ed=2724.5 pmi=0.195
        words=feature,barely,mix,looks,turning,growing,boasts,colour,survive,hardly
Topic 3 p=3.28% ew=282.2 ed=967.3 pmi=0.962
        words=voters,candidates,coalition,politician,politics,socialists,polls,elected,elections,votes
Topic 4 p=2.22% ew=404.2 ed=2425.7 pmi=0.267
        words=thus,agree,opportunity,discussions,seeks,favour,reviewed,consider,fixed,rules
Topic 5 p=2.01% ew=323.5 ed=792.6 pmi=0.285
        words=plc,newsdesk,171,542,+44,stake,digital,company,shareholders,murdoch

...

Topic root words=last,first,year,told,years,since,made,three,time,world
Topical words=xiaoping,kremlin,yeltsin,o.j,exhibitor,chirac,debakey,akchurin,quintuple,yastrzhembsky

Average topicXword sparsity = 93.56%
Average docXtopic sparsity = 92.32%
Average PMI = 0.572
conc. = 0.000000, empty = 0, exp.ent = 56.789223
```

Algorithm Design

- may use hypothesis testing at decision points inside the algorithm
  - e.g. is ethnicity relevant to risk of readmission?
- may fit the model by minimising a cost function
  - e.g. sum of squared errors between model and actual
- may fit the model by maximising likelihood
  - e.g. this is a precise recipe; given the model and data the MLE is well-defined
- what sort of bias and variance does the model have?
  - don't know the truth, but we can estimate, explore, ...

Outline

Revision Point Estimator

Revision Central Limit Theorem

Confidence Intervals

Confidence Intervals for Normal Means

Confidence Intervals for Difference of Normal Means

Approximate CIs for Sample Means

Major Questions (OPTIONAL)

Confidence Interval Normal Variance(OPTIONAL)

Using Chi-Squared Distribution

σ^2 for a Gaussian: If $x_1, ..., x_n$ is a sample from a Gaussian population with mean μ and variance σ^2. Let $\hat{\sigma}_u^2$ be the sample variance. Then $(n-1)\hat{\sigma}_u^2/\sigma^2$ is chi-squared with $n-1$ degrees of freedom.

- we know $n, \hat{\sigma}_u^2$
- we can get "reasonable" ranges of χ^2 from tables
- use the equation $\sigma^2 = \frac{n-1}{\chi^2}\hat{\sigma}_u^2$ to infer "reasonable" ranges of σ^2
- let $\chi^2_{\beta,k}$ denote the $\beta \times 100$ percentile for chi-squared with k degrees of freedom

The $((1-\alpha) \times 100)\%$ confidence interval for σ^2 is $\left[\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}\hat{\sigma}_u^2, \frac{n-1}{\chi^2_{\alpha/2,n-1}}\hat{\sigma}_u^2\right]$.

CI for unknown μ and σ^2

- Chi-squared distribution gives CI for σ^2 of

$$\left(\frac{n-1}{\chi^2_{1-\alpha/2,n-1}} \hat{\sigma}^2_u, \frac{n-1}{\chi^2_{\alpha/2,n-1}} \hat{\sigma}^2_u \right)$$

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

- Here, $\chi^2_{1-\alpha/2,n-1}$ is the $100\alpha/2$-th percentile of the chi-squared distribution with $n-1$ degrees of freedom.
- $\chi^2_{\alpha/2,n-1}$ is the $100(1-\alpha/2)$-th percentile of the chi-squared distribution with $n-1$ degrees of freedom
- Find values in *NIST critical values of the chi-square distribution*

Worked Example

- Let us revisit our Pima BMI data:

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

- We do not have access to the population variance
- Our unbiased estimate of the population variance from the sample is:

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

- Compare this to the "known variance" value of 43.75.

Worked Example, cont.

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

$$\begin{aligned} \texttt{qchisq}(\texttt{p} = 0.05/2, \texttt{df} = 7) &\approx 1.690 \\ \texttt{qchisq}(\texttt{p} = 1 - 0.05/2, \texttt{df} = 7) &\approx 16.013 \end{aligned}$$

- This results in the 95% CI for σ^2 of

$$\left(\frac{7}{16.013} 51.37, \frac{7}{1.690} 51.37 \right) = (22.46, 212.78)$$

- For σ this becomes $(4.739, 14.587)$ and the "known standard deviation" of 6.614

End of Week 6