FIT2086 Lecture 3
Parameter Estimation

Daniel F. Schmidt

Faculty of Information Technology, Monash University

August 5, 2017

Outline

Revision from Lecture 2 (1)

- $\mathbb{P}(X = x, Y = y)$ is joint probability of $X = x$ and $Y = y$.
  - Sum-rule (marginal probability):

$$\mathbb{P}(X = x) = \sum_{y} \mathbb{P}(X = x, Y = y)$$

  - Conditional probability

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

- Cumulative distribution function (for ordered x):

$$\mathbb{P}(X \leq x) = \sum_{x \leq x} \mathbb{P}(X = x)$$

- Also: $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$.

Revision from Lecture 2 (2)

- Let $\mathbb{P}(X = x) \equiv p(x)$; expectation and variance of $f(X)$:

$$\begin{aligned} \mathbb{E}\left[f(X)\right] &= \sum_x p(x) f(x) \\ \mathbb{V}\left[f(X)\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[f(X)\right]\right)^2\right] \end{aligned}$$

  with integral replacing sum for continuous RVs.

- Some useful rules:
  - $\mathbb{E}\left[f(X) + g(Y)\right] = \mathbb{E}\left[f(X)\right] + \mathbb{E}\left[g(Y)\right]$
  - $\mathbb{E}\left[cf(X)\right] = c\mathbb{E}\left[f(X)\right]$
  - $\mathbb{V}\left[cf(X)\right] = c^2\mathbb{V}\left[f(X)\right]$

- If X, Y are independent RVs
  - $\mathbb{E}\left[f(X)g(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right]$
  - $\mathbb{V}\left[f(X) + g(Y)\right] = \mathbb{V}\left[f(X)\right] + \mathbb{V}\left[g(Y)\right]$

Revision from Lecture 2 (3)

- Normal distribution; $X \in \mathbb{R}$, $X \sim N(\mu, \sigma^2)$

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

- Bernoulli distribution; $X \in \{0, 1\}$, $X \sim \text{Be}(\theta)$

$$\mathbb{E}[X] = \theta, \quad \mathbb{V}[X] = \theta(1 - \theta)$$

- Binomial distribution; $X \in \{0, 1, \ldots, n\}$, $X \sim \text{Bin}(\theta, n)$

$$\mathbb{E}[X] = n\theta, \quad \mathbb{V}[X] = n\theta(1 - \theta)$$

- Poisson distribution; $X \in \{0, 1, 2, \ldots\}$, $X \sim \text{Poi}(\lambda)$

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

Outline

Problem Statement

- Imagine we have observed some data $\mathbf{y} = (y_1, \ldots, y_n)$
  $\Rightarrow \mathbf{y}$ commonly used to denote data
- For example, heights of people in a classroom

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72)$$

- We would like to model these using a normal distribution

$$p(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(y-\mu)^2}{\sigma^2}\right)$$

  but of course, the *population* μ and σ^2 are unknown.

- Estimation: How to use the *data* to select values of μ and σ^2

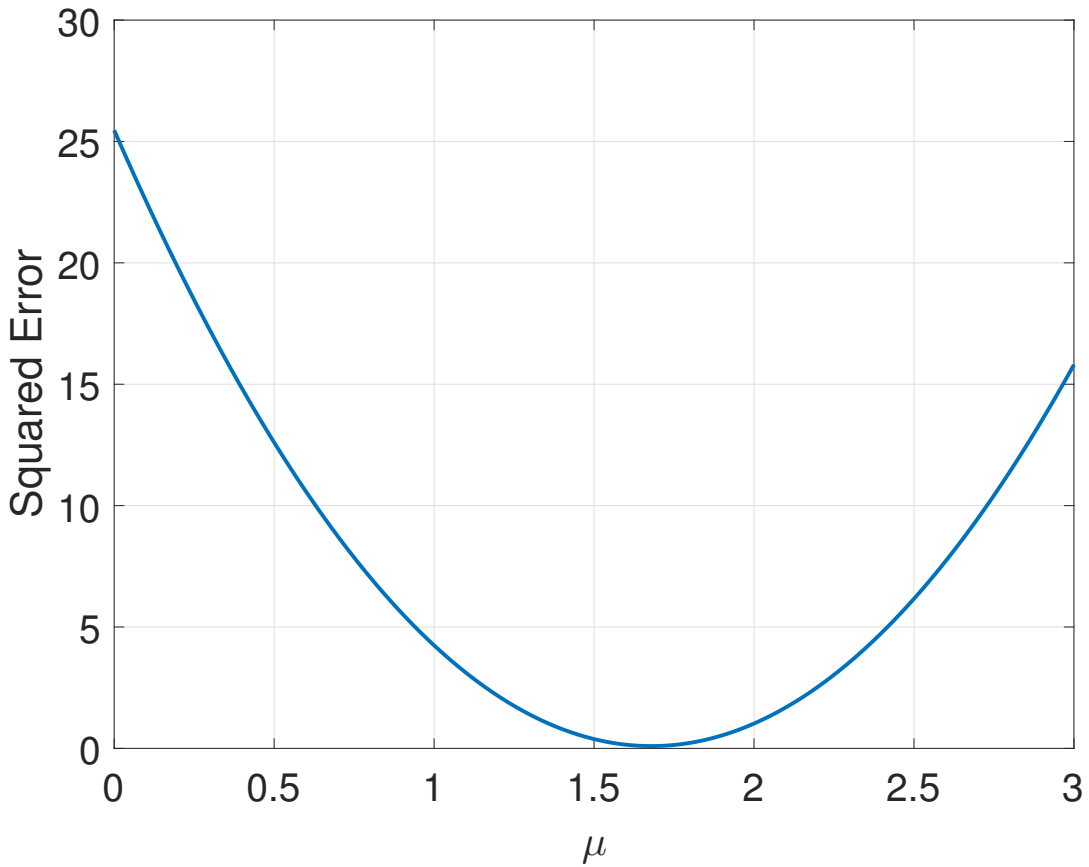
Minimum Sum of Squared Errors (1)

- Let's focus first on the mean, μ
  $\Rightarrow$ This represents the centre of the normal distribution
- One heuristic approach might be to choose a μ that is close to all the data points
- How do we measure closeness?
  $\Rightarrow$ A mathematically convenient measure is squared error:

$$\text{sse}(\mu) = \sum_{i=1}^{n} (y_i - \mu)^2$$

- Squared-error is obviously always greater than zero
- To estimate μ using this approach: adjust μ until $\text{sse}(\mu)$ attains its minimum

Minimum Sum of Squared Errors (2)



Sum of squared errors (sse) as a function of the parameter μ for our example data set. There is one clear minimum.

Minimum Sum of Squared Errors (3)

- Formally we can write this process as

$$\hat{\mu} = \arg\min_{\mu} \left\{ \sum_{i=1}^{n} (y_i - \mu)^2 \right\},$$

  where
  - $\arg\min_{x} \{f(x)\}$ means find the value of x that minimises $f(x)$;
  - $\hat{\mu}$ (read as "hat μ") denotes our estimator.

- Estimators are usually identified by putting a hat on the quantity we are estimating

Minimum Sum of Squared Errors (4)

- Due to choice of squared error, estimate is easy to find:
  1. First, differentiate $\mathrm{sse}(\mu)$ with respect to μ

$$\begin{aligned}\frac{d\,\mathrm{sse}(\mu)}{d\mu} &= \sum_{i=1}^{n} \frac{d}{d\mu}(y_i - \mu)^2 \\ &= -2\sum_{i=1}^{n}(y_i - \mu) \\ &= -2\sum_{i=1}^{n} y_i + 2n\mu\end{aligned}$$

  2. Then set the derivative to zero, and solve for μ, yielding:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

which is readily identified as the sample mean.

Minimum Squared Error (5)

- Recall our example data set:

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72)$$

- In this case, $\hat{\mu} = 1.6789$ if we minimise squared error
- But the normal distribution has two parameters: μ and σ ...
  $\Rightarrow$ How do we estimate σ?
- The minimum error approach offers no obvious measure of goodness-of-fit for σ
- A more general approach is required

Maximum Likelihood

- One solution: maximum likelihood.
- This is a very general procedure for estimating parameters of statistical models
  - Was first proposed in the 1920s by Ronald Fisher (1890–1962)
  - Heuristic procedure that has been shown to have many strong properties
  - Widely applicable to many models

Maximum Likelihood (2)

- Let us consider a probability model with parameter(s) $\boldsymbol{\theta}$.
- We measure the goodness-of-fit of a model to data by the probability it assigns to the data, i.e.,

$$p(\mathbf{y} \,|\, \boldsymbol{\theta})$$

- The larger the probability, the more likely the observed data would be under that model
- For many models we will examine, the probabilities of $y_1, y_2, \ldots, y_n$ are independent so that:

$$p(\mathbf{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \,|\, \boldsymbol{\theta})$$

Maximum Likelihood (3)

The Method of Maximum Likelihood

The method of maximum likelihood says we should use the model that assigns the greatest probability to the data we have observed.

Formally, the maximum likelihood (ML) estimator is found by solving

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{p(\mathbf{y} \mid \boldsymbol{\theta})\}$$

where $p(\mathbf{y} \mid \boldsymbol{\theta})$ is called the likelihood function.

Maximum Likelihood (4)

- In practice it is mathematically easier to solve the equivalent problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \{-\log p(\mathbf{y} \mid \boldsymbol{\theta})\}$$

  where

$$-\log(\mathbf{y} \mid \boldsymbol{\theta})$$

  is known as the negative log-likelihood.

- Sometimes we use $L(\mathbf{y} \mid \boldsymbol{\theta})$ to denote the negative log-likelihood
- Sometimes, the log-likelihood is used instead.

ML Estimation of Normal (1)

- Let's return to the problem of estimating μ and σ for a normal distribution
- For the normal distribution $\boldsymbol{\theta} = (\mu, \sigma)$.
- Given data $\mathbf{y} = (y_1, \ldots, y_n)$ the likelihood is

$$p(\mathbf{y} \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu - y_i)^2\right)$$

by

  - the independence of $y_1, \ldots, y_n$;
  - the fact that $e^{-a}e^{-b} = e^{-a-b}$.

ML Estimation of Normal (2)

- The negative log-likelihood function is then:

$$\begin{aligned} L(\mathbf{y} \mid \mu, \sigma) &= -\log p(\mathbf{y} \mid \mu, \sigma) \\ &= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \qquad (1) \end{aligned}$$

- To minimise this for μ and σ we need to differentiate equation (1) with respect to μ and σ and find the values that set the (partial) derivatives to zero, i.e., we need to solve the simultaneous equations:

$$\begin{aligned} \partial L(\mathbf{y}|\mu,\sigma)/\partial\mu &= 0, \\ \partial L(\mathbf{y}|\mu,\sigma)/\partial\sigma &= 0. \end{aligned}$$

- It turns out for this problem, this is actually quite easy

ML Estimation of Normal (2)

- The negative log-likelihood function is then:

$$\begin{aligned} L(\mathbf{y} \mid \mu, \sigma) &= -\log p(\mathbf{y} \mid \mu, \sigma) \\ &= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \qquad (1) \end{aligned}$$

- To minimise this for μ and σ we need to differentiate equation (1) with respect to μ and σ and find the values that set the (partial) derivatives to zero, i.e., we need to solve the simultaneous equations:

$$\begin{aligned} \partial L(\mathbf{y}|\mu,\sigma)/\partial\mu &= 0, \\ \partial L(\mathbf{y}|\mu,\sigma)/\partial\sigma &= 0. \end{aligned}$$

- It turns out for this problem, this is actually quite easy

ML Estimation of Normal (3)

- Partial derivative with respect to μ:

$$\begin{aligned} \frac{\partial L(\mathbf{y} \mid \mu, \sigma)}{\partial \mu} &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \\ &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} y_i + \frac{n\mu}{\sigma^2} \end{aligned} \tag{2}$$

  which is similar to our minimum squared error estimator.

- In fact, setting equation (3) to zero and solving for μ yields

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

  which is again, just the sample mean.

ML Estimation of Normal (4)

- However, ML also gives us a clear recipe for estimating σ
- Plugging $\hat{\mu}$ into $L(\mathbf{y}|\mu,\sigma)$ removes μ from the equation
- Partial derivative with respect to σ:

$$\begin{aligned}\frac{\partial L(\mathbf{y}\,|\,\hat{\mu},\sigma)}{\partial\sigma} &= \frac{\partial}{\partial\sigma}\frac{n}{2}\left[\log\sigma^2+\log(2\pi)\right]+\sum_{i=1}^{n}(y_i-\hat{\mu})^2\frac{\partial}{\partial\sigma}\frac{1}{2\sigma^2}\\ &= \frac{n}{\sigma}-\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \qquad (3)\end{aligned}$$

where we use the facts that

- $\log(ab)=\log b+\log b$;
- $\frac{\partial}{\partial x}Kf(z)f(x)=Kf(z)\frac{\partial}{\partial x}f(x)$; and
- $\frac{\partial}{\partial x}f(z)=0$.

ML Estimation of Normal (5)

- Solving for σ:

$$\frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 0$$

$$\Rightarrow \quad \sigma^3\left[\frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \hat{\mu})^2\right] = 0$$

$$\Rightarrow \quad n\sigma^2 - \sum_{i=1}^{n}(y_i - \hat{\mu})^2 = 0$$

$$\Rightarrow \quad n\sigma^2 = \sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

$$\Rightarrow \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

ML Estimation of Normal (5)

- The ML estimator for σ is:

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

  which can be identified as the sample standard deviation.

- So, for the normal distribution, the ML estimators are:
  - The sample mean for μ;
  - The sample standard deviation for σ

Normal Example (1)

- Recall our example data set:

$$\mathbf{y} = (1.75, 1.64, 1.81, 1.55, 1.51, 1.67, 1.83, 1.63, 1.72)$$

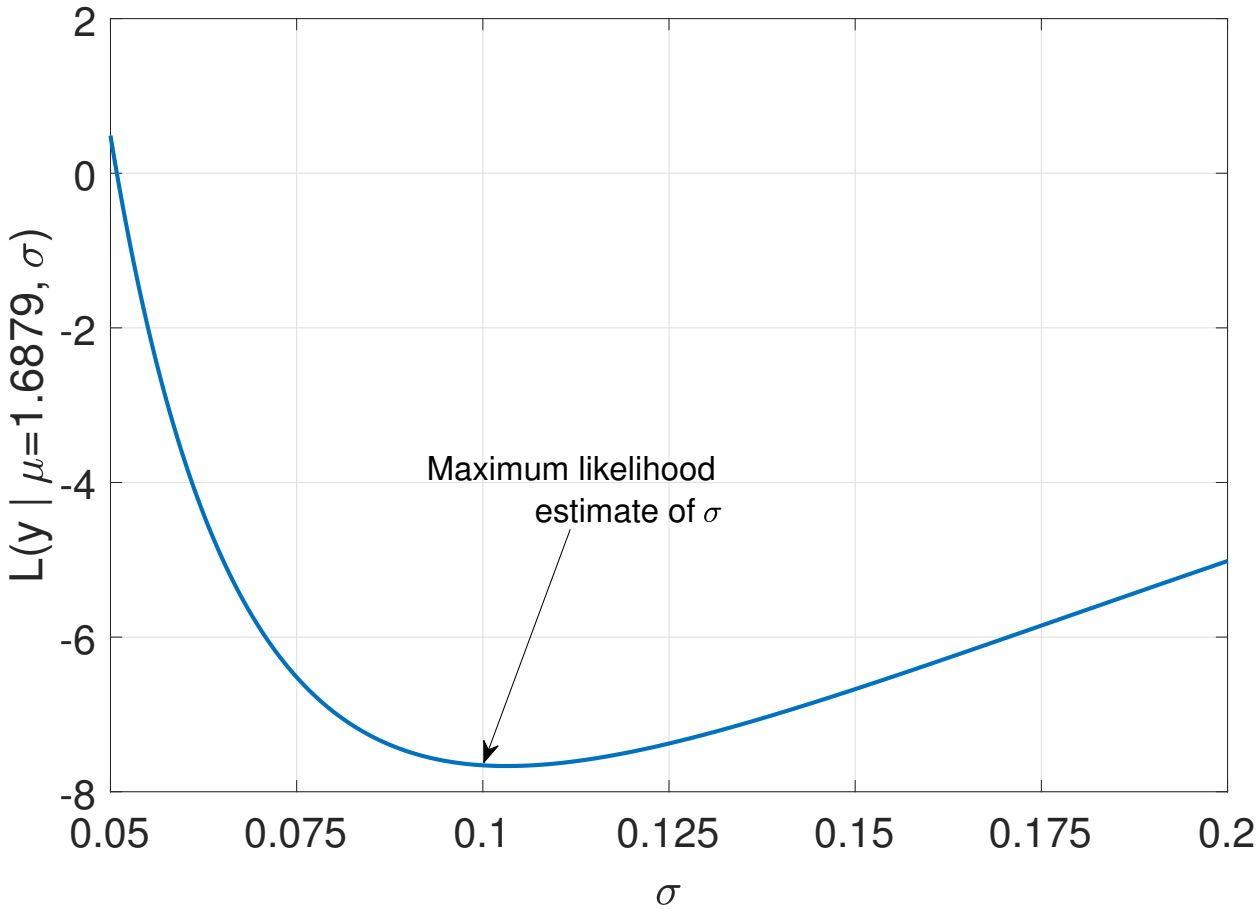
- Using maximum likelihood to fit a normal distribution to this data, we have:

$$\hat{\mu} = 1.6789$$

and

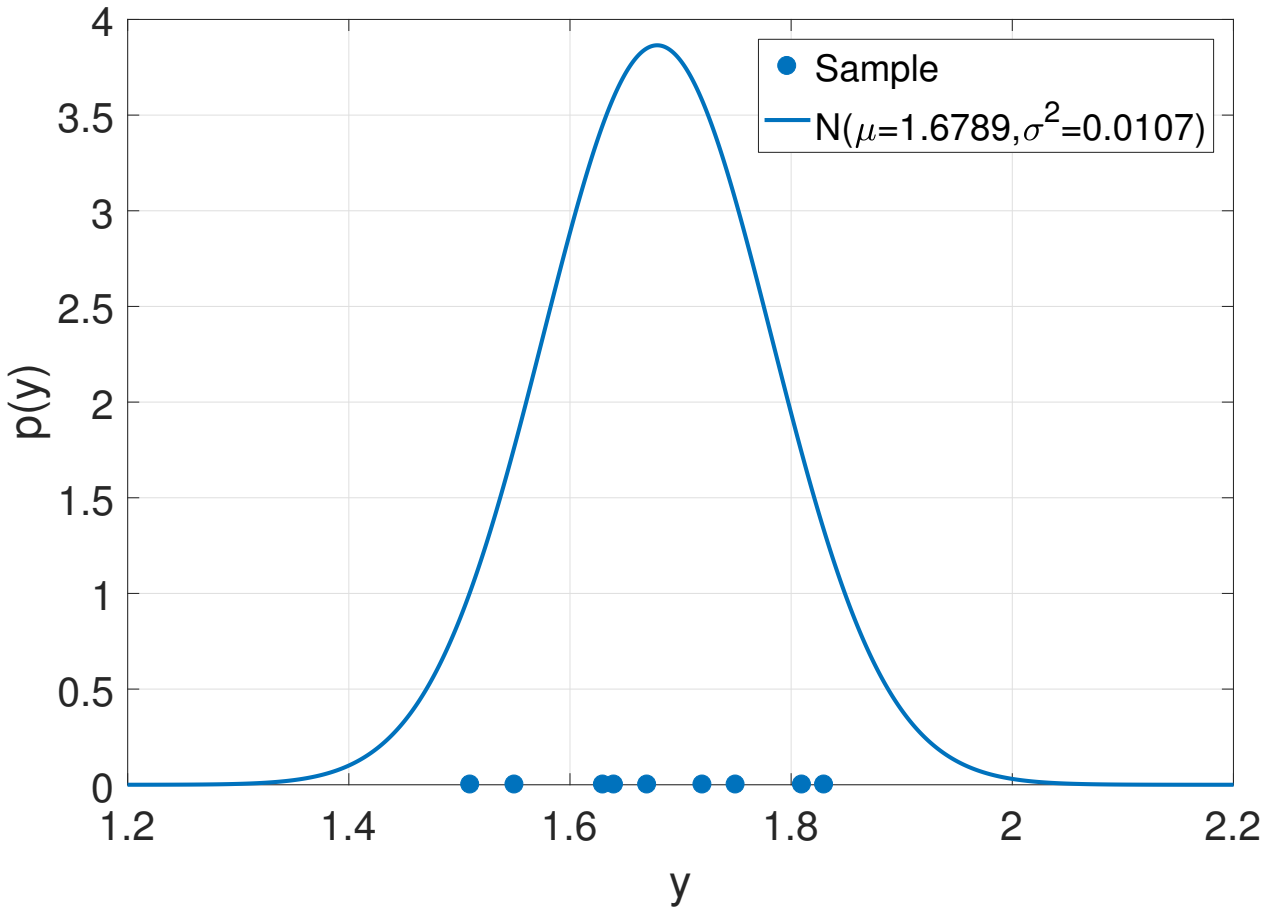
$$\begin{aligned} \hat{\sigma} &= \sqrt{\frac{1}{9}\sum_{i=1}^{9}(y_i - 1.6789)^2} \\ &= 0.1032 \end{aligned}$$

Normal Example (2)



Negative log-likelihood $L(\mathbf{y} \mid \mu = \hat{\mu}, \sigma)$ as a function of σ with μ fixed at the maximum likelihood estimate $\hat{\mu} = 1.6789$. Samples were $\mathbf{y} = (7, 9, 3, 5, 3, 4, 4, 9, 4, 6)$.

Normal Example (3)



Data samples and the normal distribution fitted by maximum likelihood with $\hat{\mu} = 1.6879$ and $\hat{\sigma} = 0.1032$. Note that the bulk of the samples lie within $(\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}) \approx (1.47, 1.88)$

How can we use our estimates?

- We can use these estimates to make statements/predictions about the population
- To do this, we use them in the distribution, i.e., $p(y \mid \hat{\mu}, \hat{\sigma}^2)$
  $\Rightarrow$ this is called the "plug-in" distribution
- Can use plug-in distribution to make probability statements
- In our example, we could ask "what is the probability a person from our population has a height between $1.6m$ and $1.8m$"?, which is estimated by

$$\mathbb{P}(1.6 < X < 1.8 \mid \hat{\mu} = 1.6879, \hat{\sigma}^2 = 0.1032) \approx 0.664$$

$\Rightarrow$ The better our estimates, the more accurate the answers

Properties of ML

- The original maximum likelihood proposal was heuristic
- But a large body of research since has shown ML has many good properties
- We will briefly touch on a couple of important ones:
  - Parameterization invariance
  - Statistical consistency

Parameterization Invariance (1)

- Parameters of a model are just labels
  $\Rightarrow$ they specify a distribution from a collection of distributions
- This means it is possible to reparameterise any distribution in terms of new parameters (new labels)
- For example, in the normal case we could use the mean μ and inverse-variance (precision) $\phi = 1/\sigma^2$:

$$p(y \mid \mu, \phi) = \left(\frac{\phi}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\phi(y-\mu)^2}{2}\right).$$

- This parametric model indexes exactly the same set of distributions, but using different labels
  $\Rightarrow N(\mu = 2, \sigma^2 = 4)$ is the same as $N(\mu = 2, \phi = 1/4)$.

Parameterization Invariance (1)

- Parameters of a model are just labels
  $\Rightarrow$ they specify a distribution from a collection of distributions
- This means it is possible to reparameterise any distribution in terms of new parameters (new labels)
- For example, in the normal case we could use the mean μ and inverse-variance (precision) $\phi = 1/\sigma^2$:

$$p(y \mid \mu, \phi) = \left(\frac{\phi}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\phi(y-\mu)^2}{2}\right).$$

- This parametric model indexes exactly the same set of distributions, but using different labels
  $\Rightarrow N(\mu = 2, \sigma^2 = 4)$ is the same as $N(\mu = 2, \phi = 1/4)$.

Parameterization Invariance (2)

- Parameterization invariance means that the plug-in distribution chosen by an estimator is the same regardless of the parameterization used
- For example, let us imagine we estimate the mean μ and precision $\phi = 1/\sigma^2$ using maximum likelihood.
- Parameterization invariance would imply that our ML estimate of standard deviation, $\hat{\sigma}$, we previously derived would satisfy:

$$\hat{\sigma} = \sqrt{\frac{1}{\phi}}.$$

- It is not hard to show that this is the case.

Parameterization Invariance (3)

Parameterization Invariance of Maximum Likelihood

Formally, if $p(\mathbf{y} \mid \theta)$ and $p(\mathbf{y} \mid \phi = f(\theta))$ are the same model, then

$$\hat{\theta} = f^{-1}(\hat{\phi})$$

if $\hat{\theta}$ and $\hat{\phi}$ are maximum likelihood estimators, and $f^{-1}(\cdot)$ is the inverse transform from ϕ to θ.

The above only holds if $\phi = f(\theta)$ is a one-to-one transformation; that is for every θ, there exists one (and only one) ϕ.

Parameterization Invariance (4)

- To see that ML is parameterization invariant is not difficult
- It comes from the fact that the measure of "goodness-of-fit" of a model used by ML,

$$p(\mathbf{y} \mid \boldsymbol{\theta}),$$

  is not directly based on the values of the parameters, but rather on the distribution they index
- The probability assigned to data $\mathbf{y}$ depends on the distribution of probabilities, not on the way the parameters are chosen to represent the distributions.
- Not all estimators have this important property.

ML Estimation of a Poisson (1)

- Let us look at another example of ML estimation
- Recall the Poisson distribution with rate λ:

$$p(y \mid \lambda) = \frac{\lambda \exp(-y\lambda)}{y!}.$$

- If $\mathbf{y} = (y_1, \ldots, y_n)$ are n integers, then the likelihood for a Poisson model is

$$p(\mathbf{y} \mid \lambda) = \prod_{i=1}^{n} p(y_i \mid \lambda) = \frac{\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)}{\prod_{i=1}^{n} y_i!}$$

by independence of $y_1, \ldots, y_n$.

ML Estimation of a Poisson (2)

- The negative log-likelihood is then

$$L(\mathbf{y} \mid \lambda) = -n \log \lambda + \lambda \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \log y_i!$$

- To find the ML estimator of λ we need to minimise $L(\mathbf{y} \mid \lambda)$, or equivalently solve

$$\frac{\partial L(\mathbf{y} \mid \lambda)}{\partial \lambda} = 0,$$

for λ.

ML Estimation of a Poisson (3)

- The derivative is given by

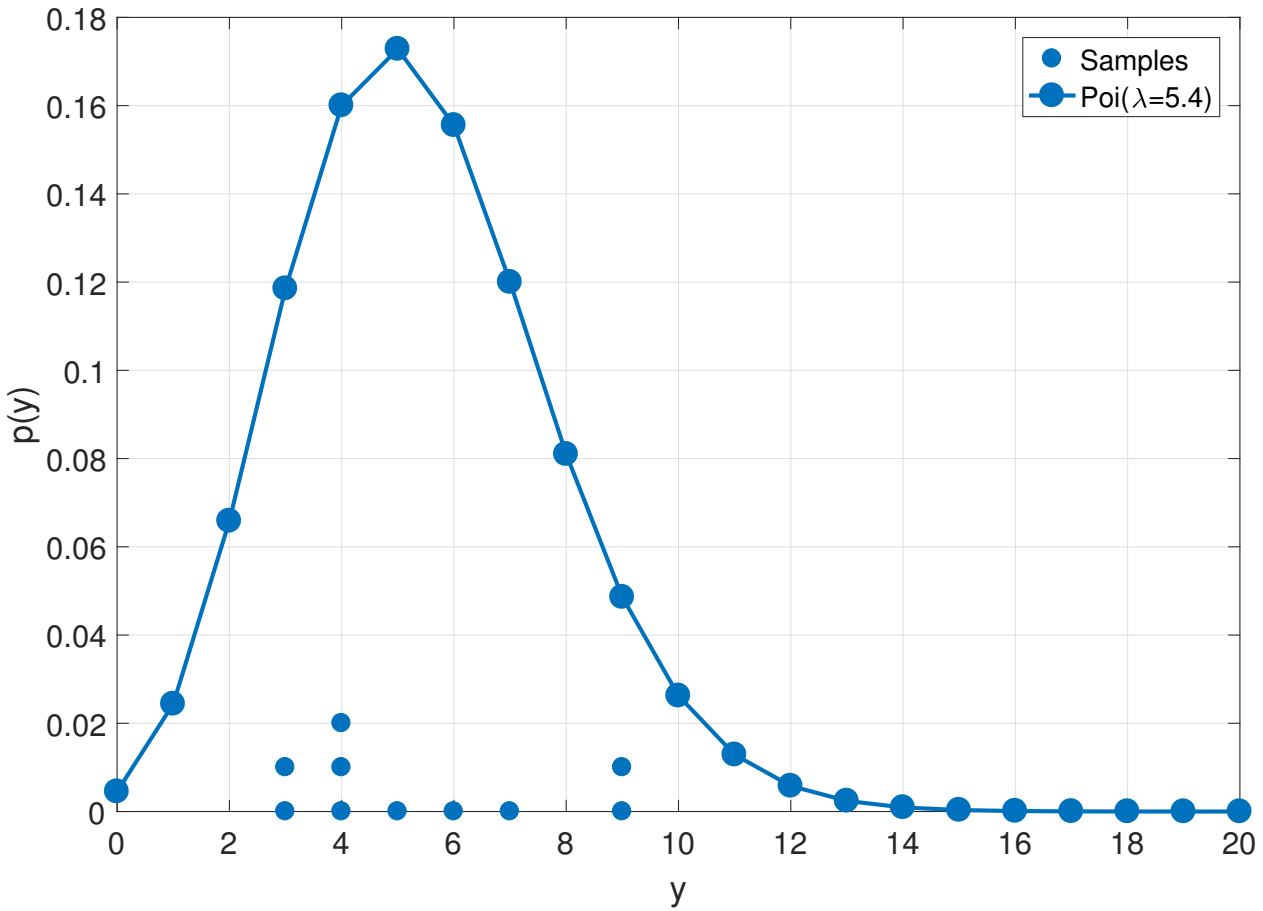
$$\frac{dL(\mathbf{y} \mid \lambda)}{d\lambda} = -\frac{n}{\lambda} + \sum_{i=1}^{n} y_i \tag{4}$$

- Setting (4) to zero and solving for λ yields

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

- So, the ML estimator for the Poisson rate is the sample mean.
  $\Rightarrow$ this is not always the case! :)

Poisson Example



Data samples and the Poisson distribution fitted by maximum likelihood with $\hat{\lambda} = 5.4$. Samples were $\mathbf{y} = (7, 9, 3, 5, 3, 4, 4, 9, 4, 6)$.

Outline

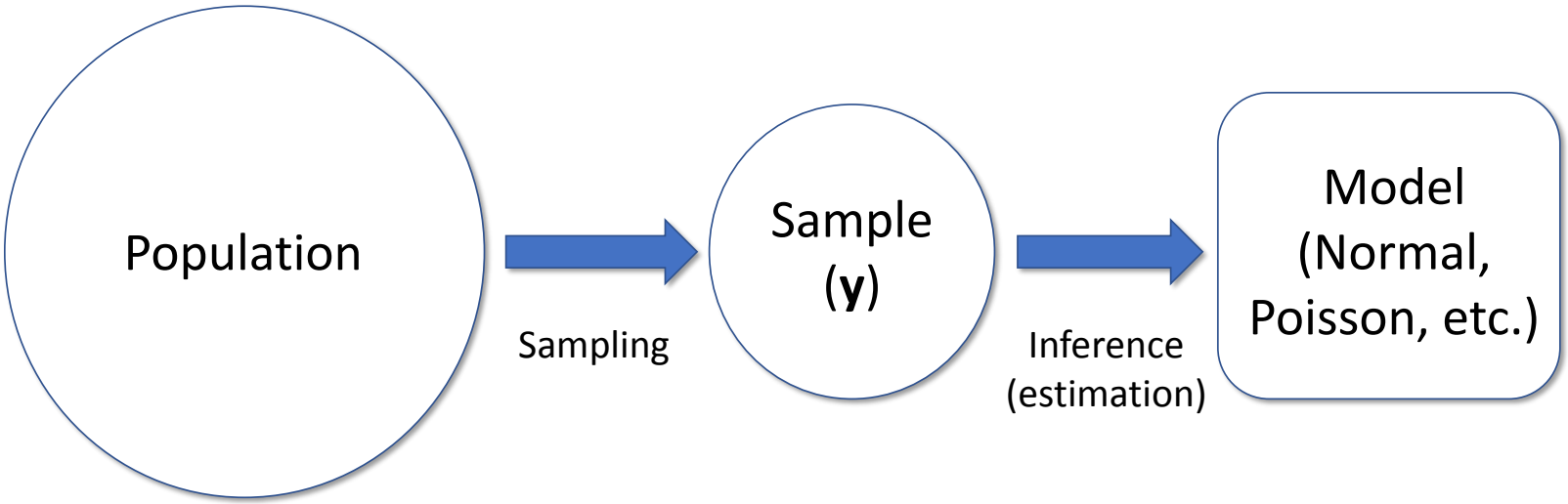
Comparing Estimators

- There are other estimators out there beyond ML
- For example, the unbiased estimator of variance:

$$\hat{\sigma}_U^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

- How to compare estimators?
  $\Rightarrow$ We can look at the different properties they have, and how accurate they are
- Many of these properties require the idea of sampling statistics.
- That is, how do the estimators behave under repeated sampling of data from the population

From population to sample to models



- The data sample $\mathbf{y}$ we that we have observed is just *one of infinitely many* different datasets we could have observed
- By taking a new random sample from the population we end up with a different dataset

Sampling Distribution of the Mean (1)

- Let's assume the following:
  1. We are measuring the heights of people in a population (in m)
  2. The population heights follow a $N(\mu = 1.65, \sigma^2 = 0.1)$ distribution
  3. We are drawing a sample of $n = 5$ measurements
  4. We are interested in the mean height of people in our population

- Recall the sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

with $\mathbf{y} = (y_1, \ldots, y_n)$ a sample from a population.

Sampling Distribution of the Mean (2)

- Imagine our sample we obtained was

$$\mathbf{y} = (1.620, 1.652, 1.623, 1.475, 1.621)$$

- Then the ML estimate of the mean would be $\bar{y} = 1.598$.
- But what if we repeated the sampling?
  $\Rightarrow$ would get a different sample, with a different sample mean
- How much different? How variable?
  $\Rightarrow$ and how accurate is our estimate?

Sampling Distribution of the Mean (2)

- Under repeated sampling from the population we can take many different samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ each of size $n = 5$
- For our example, we might have:

$$\mathbf{y}^{(1)} = (1.620, 1.652, 1.623, 1.475, 1.621) \quad \Rightarrow \bar{y}_1 = 1.598$$
$$\mathbf{y}^{(2)} = (1.729, 1.517, 1.417, 1.505, 1.683) \quad \Rightarrow \bar{y}_2 = 1.570$$
$$\mathbf{y}^{(3)} = (1.689, 1.695, 1.637, 1.668, 1.602) \quad \Rightarrow \bar{y}_3 = 1.658$$
$$\mathbf{y}^{(4)} = (1.736, 1.513, 1.695, 1.565, 1.616) \quad \Rightarrow \bar{y}_4 = 1.625$$
$$\mathbf{y}^{(5)} = (1.705, 1.753, 1.538, 1.776, 1.716) \quad \Rightarrow \bar{y}_5 = 1.697$$
$$\vdots$$

- Each one has a different sample mean

Sampling Distribution of the Mean (2)

- Under repeated sampling from the population we can take many different samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ each of size $n = 5$
- For our example, we might have:

$$\begin{aligned}
\mathbf{y}^{(1)} &= (1.620, 1.652, 1.623, 1.475, 1.621) && \Rightarrow \bar{y}_1 = 1.598 \\
\mathbf{y}^{(2)} &= (1.729, 1.517, 1.417, 1.505, 1.683) && \Rightarrow \bar{y}_2 = 1.570 \\
\mathbf{y}^{(3)} &= (1.689, 1.695, 1.637, 1.668, 1.602) && \Rightarrow \bar{y}_3 = 1.658 \\
\mathbf{y}^{(4)} &= (1.736, 1.513, 1.695, 1.565, 1.616) && \Rightarrow \bar{y}_4 = 1.625 \\
\mathbf{y}^{(5)} &= (1.705, 1.753, 1.538, 1.776, 1.716) && \Rightarrow \bar{y}_5 = 1.697 \\
&\vdots
\end{aligned}$$

- Each one has a different sample mean

Sampling Distribution of the Mean (2)

- Under repeated sampling from the population we can take many different samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ each of size $n = 5$
- For our example, we might have:

$$\begin{aligned}
\mathbf{y}^{(1)} &= (1.620, 1.652, 1.623, 1.475, 1.621) && \Rightarrow \bar{y}_1 = 1.598 \\
\mathbf{y}^{(2)} &= (1.729, 1.517, 1.417, 1.505, 1.683) && \Rightarrow \bar{y}_2 = 1.570 \\
\mathbf{y}^{(3)} &= (1.689, 1.695, 1.637, 1.668, 1.602) && \Rightarrow \bar{y}_3 = 1.658 \\
\mathbf{y}^{(4)} &= (1.736, 1.513, 1.695, 1.565, 1.616) && \Rightarrow \bar{y}_4 = 1.625 \\
\mathbf{y}^{(5)} &= (1.705, 1.753, 1.538, 1.776, 1.716) && \Rightarrow \bar{y}_5 = 1.697 \\
&\vdots
\end{aligned}$$

- Each one has a different sample mean

Sampling Distribution of the Mean (2)

- Under repeated sampling from the population we can take many different samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ each of size $n = 5$
- For our example, we might have:

$$\begin{aligned}
\mathbf{y}^{(1)} &= (1.620, 1.652, 1.623, 1.475, 1.621) && \Rightarrow \bar{y}_1 = 1.598 \\
\mathbf{y}^{(2)} &= (1.729, 1.517, 1.417, 1.505, 1.683) && \Rightarrow \bar{y}_2 = 1.570 \\
\mathbf{y}^{(3)} &= (1.689, 1.695, 1.637, 1.668, 1.602) && \Rightarrow \bar{y}_3 = 1.658 \\
\mathbf{y}^{(4)} &= (1.736, 1.513, 1.695, 1.565, 1.616) && \Rightarrow \bar{y}_4 = 1.625 \\
\mathbf{y}^{(5)} &= (1.705, 1.753, 1.538, 1.776, 1.716) && \Rightarrow \bar{y}_5 = 1.697 \\
&\vdots
\end{aligned}$$

- Each one has a different sample mean

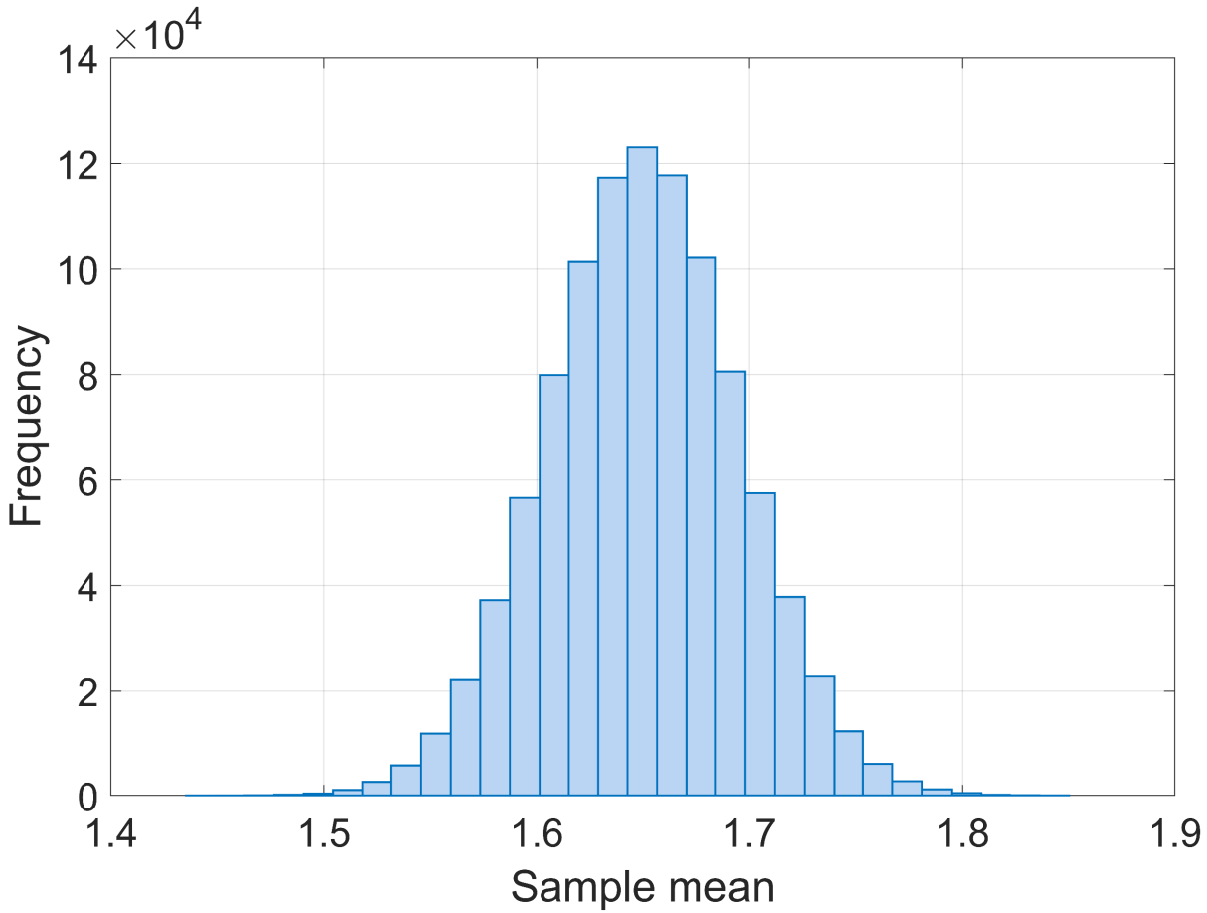
Sampling Distribution of the Mean (2)

- Under repeated sampling from the population we can take many different samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ each of size $n = 5$
- For our example, we might have:

$$\begin{aligned}
\mathbf{y}^{(1)} &= (1.620, 1.652, 1.623, 1.475, 1.621) && \Rightarrow \bar{y}_1 = 1.598 \\
\mathbf{y}^{(2)} &= (1.729, 1.517, 1.417, 1.505, 1.683) && \Rightarrow \bar{y}_2 = 1.570 \\
\mathbf{y}^{(3)} &= (1.689, 1.695, 1.637, 1.668, 1.602) && \Rightarrow \bar{y}_3 = 1.658 \\
\mathbf{y}^{(4)} &= (1.736, 1.513, 1.695, 1.565, 1.616) && \Rightarrow \bar{y}_4 = 1.625 \\
\mathbf{y}^{(5)} &= (1.705, 1.753, 1.538, 1.776, 1.716) && \Rightarrow \bar{y}_5 = 1.697 \\
&\vdots
\end{aligned}$$

- Each one has a different sample mean

Sampling Distribution of the Mean (3)



Histogram of sample means of $1,000,000$ different data samples, each of size $n = 5$, generated from a $N(\mu = 1.65, \sigma = 0.1)$ distribution.

Sampling Distributions (1)

- This slide presents the key idea behind sampling distributions:
  1. If the data sample $\mathbf{y}$ is a *realisation* of n random variables $(Y_1, \ldots, Y_n)$ from some *population* distribution,
  2. then any function of $\mathbf{y}$ is *also* a realisation of a random variable,
  3. and therefore follows a distribution of its own, which is based on the distribution of $(Y_1, \ldots, Y_n)$.
- Formally, an estimator is just a function from the sample $\mathbf{y}$ to the parameter space.
- So therefore, the values of an estimator follow a distribution based on the population distribution
- To find this distribution we need to make some assumptions about the population distribution

Sampling Distributions (2)

- A common assumption is that the population distribution is itself one of the standard parametric distributions (i.e., normal, Poisson, etc.)
  - There are ways of weakening these assumptions
  - But they also weaken the statements we can make about the sampling distribution
- Assume $Y_1, \ldots, Y_n$ follow a parametric distribution $p(\mathbf{y} \mid \boldsymbol{\theta})$
  $\Rightarrow$ in this instance we call $\boldsymbol{\theta}$ the *population* parameters
- An estimator $\hat{\boldsymbol{\theta}}(Y_1, \ldots, Y_n)$ is a function of $Y_1, \ldots, Y_n$, so it follows that

$$\hat{\boldsymbol{\theta}} \sim P(\boldsymbol{\theta})$$

- The estimator follows a distribution determined by $p(\mathbf{y} \mid \boldsymbol{\theta})$

Example: Sampling Distribution of the Mean (1)

- As an example, consider our previous example:
  - Our population is described by $Y \sim N(\mu, \sigma^2)$
  - We draw a sample $Y_1, \ldots, Y_n$ of size n from the population
  - We estimate the population mean using the sample mean
- Recall the definition of the sample mean

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

- What is the distribution of $\bar{Y}$?

Example: Sampling Distribution of the Mean (2)

- Under our assumed population distribution we have

$$Y_i \sim N(\mu, \sigma^2)$$

for $Y_1, \ldots, Y_n$

- We can use the following facts
  1. If $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$ then

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

  2. If $Y_1 \sim N(\mu, \sigma^2)$ then

$$Y_1/n \sim N(\mu/n, (\sigma/n)^2)$$

Example: Sampling Distribution of the Mean (3)

- Note that the sample mean can be rewritten as

$$\bar{Y} = \frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}$$

Then, using the facts from the previous slide it is easy to see that:

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

- Note: the variance of the sample mean
  1. decreases with increasing sample size n;
  2. increases with increasing population variance.
- The distribution may not always be easily found analytically – but using simulation we can always get an approximation

Example: Sampling Distribution of the Mean (3)

- Note that the sample mean can be rewritten as

$$\bar{Y} = \frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}$$

  Then, using the facts from the previous slide it is easy to see that:

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

- Note: the variance of the sample mean
  1. decreases with increasing sample size n;
  2. increases with increasing population variance.
- The distribution may not always be easily found analytically – but using simulation we can always get an approximation

How can we use this 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
- We will examine the first two over the next two weeks
- For now, we will only look at the third use

Evaluating Estimators

- As was previously stated, there are many different estimators that exist beyond maximum likelihood
- Imagine we were estimating a parameter θ of a distribution
- Consider two different estimators, say $\hat{\theta}_1$ and $\hat{\theta}_2$.
- How could we compare the two and decide if one is superior to the other?
- We could ask:
  - Does one exhibit more of a systematic bias than the other?
  - Is one more variable than the other?
  - Is one closer on average to the population parameter θ?
- Let's start with the first question

Estimator Bias (1)

- Estimator bias is the degree to which an estimator tends to over/underestimates the population parameter
- Let $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)$ be our data
- If $\hat{\theta}(\underset{\sim}{Y})$ is an estimator of a parameter θ, then it's bias is

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}(\underset{\sim}{Y})\right] - \theta$$

  where the expectation is taken with respect the RVs $\underset{\sim}{Y}$
- The bias is a function of the population parameter θ
  $\Rightarrow$ bias could be worse for certain values of θ
- If $b_\theta(\hat{\mu}(\underset{\sim}{Y})) = 0$ for all θ we say the estimator is unbiased.

Estimator Bias (1)

- Estimator bias is the degree to which an estimator tends to over/underestimates the population parameter
- Let $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)$ be our data
- If $\hat{\theta}(\underset{\sim}{Y})$ is an estimator of a parameter θ, then it's bias is

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}(\underset{\sim}{Y})\right] - \theta$$

  where the expectation is taken with respect the RVs $\underset{\sim}{Y}$
- The bias is a function of the population parameter θ
  $\Rightarrow$ bias could be worse for certain values of θ
- If $b_\theta(\hat{\mu}(\underset{\sim}{Y})) = 0$ for all θ we say the estimator is unbiased.

Estimator Bias (2)

- Let $Y_1, \ldots, Y_n$ be an i.i.d. RVs with mean μ
  $\Rightarrow$ no further assumptions about population distribution

- For example, the bias of the sample mean $\bar{Y}$ is zero since

$$\begin{aligned}
\mathbb{E}\left[\bar{Y}\right] &= \mathbb{E}\left[\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right] \\
&= \frac{\mathbb{E}\left[Y_1\right]}{n} + \frac{\mathbb{E}\left[Y_2\right]}{n} + \cdots + \frac{\mathbb{E}\left[Y_n\right]}{n} \\
&= \mu
\end{aligned}$$

where $\mathbb{E}\left[Y_i\right] = \mu$ by our assumption.

$\Rightarrow$ So under weak assumptions, sample mean $\bar{Y}$ is an unbiased estimator of the population mean.

Variance of an Estimator (1)

- The variance of an estimator $\hat{\theta}(\underset{\sim}{Y})$ is defined as

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}(\underset{\sim}{Y}) - \mathbb{E}\left[\hat{\theta}(\underset{\sim}{Y})\right]\right)^2\right] = \mathbb{V}\left[\hat{\theta}(\underset{\sim}{Y})\right]$$

  where the expectation is taken with respect the RVs $\underset{\sim}{Y}$
- Again, is a function of the population parameters
- Estimator variance is equal to the variance of the *sampling distribution* of $\hat{\theta}$
- The larger the variance, the more we expected the estimates we get to vary if we resampled from the population

Variance of an Estimator (2)

- Let $Y_1, \ldots, Y_n$ be i.i.d. RVs with mean μ and variance σ^2
- The variance of the sample mean $\bar{Y}$ is then

$$\begin{aligned}
\mathbb{V}\left[\bar{Y}\right] &= \mathbb{V}\left[\frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}\right] \\
&= \mathbb{V}\left[\frac{Y_1}{n}\right] + \mathbb{V}\left[\frac{Y_2}{n}\right] + \cdots + \mathbb{V}\left[\frac{Y_n}{n}\right] \\
&= \frac{1}{n^2}\left(\mathbb{V}\left[Y_1\right] + \mathbb{V}\left[Y_2\right] + \cdots + \mathbb{V}\left[Y_n\right]\right) \\
&= \sigma^2/n
\end{aligned}$$

  where we use: (i) the independence of the RVs in step 2, (ii) the fact that $\mathbb{V}\left[kX\right] = k^2\mathbb{V}\left[X\right]$ in step 3, and (iii) the fact that $\mathbb{V}\left[Y_i\right] = \sigma^2$ (by assumption) in step 4.
- So the larger n, the less variable the sample mean becomes

Strength of Assumptions

- Let's compare the results we just derived:
  1. If $Y_1, \ldots, Y_n$ are i.i.d. RVs with mean μ and variance σ^2

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

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

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

- Both give same information re: the mean and variance of $\bar{Y}$
  $\Rightarrow$ but the latter has more information regarding distribution
- The best we could do with the former is use Chebychev's Theorem to get bounds on statements like

$$\mathbb{P}(y_1 < \bar{Y} < y_2)$$

  while we could get exact probabilities in the case of the latter – but assumptions are stronger!

Mean squared error of an Estimator (1)

- It is possible for one estimator to be more biased than another
- It is also possible for one estimator to have smaller variance than another
- How could we decide if one is better than the other?
- We could use mean squared error (MSE) of estimator:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta}(\underset{\sim}{Y}) - \theta)^2\right]$$

- Measures how far on average the estimator is from the population parameter
  $\Rightarrow$ the larger the value, the poorer the estimator

Mean squared error of an Estimator (2)

- Mean squared-error is not the only "error" measure
  $\Rightarrow$ For example, could use absolute error instead
- However, MSE has some nice mathematical properties
- Perhaps the most important is the bias-variance decomposition:

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

- That is, the mean squared error of an estimator can be broken into the sum of:
  - the square of the bias of the estimator, and
  - the variance of the estimator.
- For unbiased estimators, the mean squared error reduces to the variance of the estimator

Mean squared error of an Estimator (3)

- Let $Y_1, \ldots, Y_n$ be an i.i.d. RVs with mean μ and variance σ^2
- Then the mean squared error of the sample mean, as an estimator of the population mean μ is:

$$\mathrm{SE}_{\mu,\sigma^2}(\bar{Y}) = b_\mu^2(\bar{Y}) + \mathrm{Var}_{\mu,\sigma^2}(\bar{Y}) = \sigma^2/n$$

as the bias is zero, as was previously established.

- Three points of note:
  - The mean squared error of the sample mean, under our assumptions, is the same irrespective of μ;
  - The mean squared error is an increasing function of σ^2
  - The mean squared error is a decreasing function of the n

Mean squared error of an Estimator (3)

- Let $Y_1, \ldots, Y_n$ be an i.i.d. RVs with mean μ and variance σ^2
- Then the mean squared error of the sample mean, as an estimator of the population mean μ is:

$$\mathrm{SE}_{\mu,\sigma^2}(\bar{Y}) = b_\mu^2(\bar{Y}) + \mathrm{Var}_{\mu,\sigma^2}(\bar{Y}) = \sigma^2/n$$

  as the bias is zero, as was previously established.

- Three points of note:
  - The mean squared error of the sample mean, under our assumptions, is the same irrespective of μ;
  - The mean squared error is an increasing function of σ^2
  - The mean squared error is a decreasing function of the n

Comparing Estimators (1)

- So given these metrics, we can compare estimators.
- As example, the maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

where $\bar{y}$ is the sample mean.

- An alternative estimator of σ^2 is the unbiased estimator

$$\hat{\sigma}^2_{\mathrm{u}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\frac{n}{n-1}\right)\hat{\sigma}^2_{\mathrm{ML}}$$

which is always bigger than $\hat{\sigma}^2_{\mathrm{ML}}$.

Comparing Estimators (2)

- Let us assume $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$
- The bias of the ML estimator of σ^2 is then known to be

$$b_{\sigma^2}(\hat{\sigma}^2_{ML}) = -\frac{\sigma^2}{n}$$

while

$$b_{\sigma^2}(\hat{\sigma}^2_{\mathrm{u}}) = 0$$

for all σ^2, i.e., it is unbiased.

- However, the variance of the unbiased estimator satisfies

$$\mathrm{Var}_{\sigma^2}(\hat{\sigma}^2_{\mathrm{u}}) = \left(\frac{n}{n-1}\right)^2 \mathrm{Var}_{\sigma^2}(\hat{\sigma}^2_{\mathrm{ML}})$$

which is always bigger than $\mathrm{Var}_{\sigma^2}(\hat{\sigma}^2_{\mathrm{ML}})$.

Comparing Estimators (3)

- One weakness of bias, variance, mean squared error is that they depend on the particular choice of parameterisation
- For example, while $\hat{\sigma}_{\mathrm{u}}^2$ is an unbiased estimator of the variance, it turns out that

$$\mathbb{E}\left[\sqrt{\hat{\sigma}_{\mathrm{u}}^2}\right] - \sigma \neq 0$$

  so that it is a biased estimator of the standard deviation.
- There exist error measures to use in place of mean squared error, etc. that don't depend on the particular parameterisation we choose

Consistency (1)

- Loosely speaking, an estimator $\hat{\theta}$ is consistent if for increasing sample sizes $n \to \infty$ it gets closer and closer to the population parameter θ.
- Imples that for large enough samples, we can be sure our estimate is good.
- Consistency is not always easy to prove, but one result is useful. An estimator $\hat{\theta}$ is consistent if

$$\begin{aligned} b_\theta(\hat{\theta}) &\to 0, \\ \mathrm{Var}_\theta(\hat{\theta}) &\to 0, \end{aligned}$$

  as $n \to \infty$ for all θ.
- Consistency is a desirable property that *does not* depend on the parameterisation of the problem.

Consistency (2)

- Let $Y_1, \ldots, Y_n$ are i.i.d. RVs with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$
- Then, the sample mean $\bar{Y}$ is a consistent estimator of the population mean μ as

$$\begin{aligned} b_\mu(\bar{Y}) &= 0 \\ \mathrm{Var}_{\sigma^2}(\bar{Y}) &= \sigma^2/n \end{aligned}$$

  and clearly $\sigma^2/n \to 0$ as $n \to \infty$.
- We could also have proved this from the Weak Law of Large Numbers from last lecture.

- For many distributions, maximum likelihood is consistent
  $\Rightarrow$ there do exist some problems for which it is not

Reading/Terms to Revise

- Reading for this week: Chapters 6 (Sections 6.1,6.2,6.4,6.5) and 7 (Sections 7.1, 7.2, 7.7) of Ross.
- Terms you should know:
  - Estimator, parameter estimation
  - Sample mean
  - Maximum likelihood;
  - Sampling distribution;
  - Bias, variance, and squared error of an estimator;
- Next week we will cover the Central Limit Theorem and confidence intervals, which are both related to sampling statistics and parameter estimators.