Point estimation

(Module 2)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2018

Contents

Aims of this module

- Introduce the main elements of statistical inference and estimation, especially the idea of a sampling distribution
- Show the simplest type of estimation: that of a single number
- Show some general approaches to estimation, especially the method of maximum likelihood

1 Estimation & sampling distributions

Motivating example

On a particular street, we measure the time interval (in minutes) between each car that passes:

```
2.55  2.13  3.18  5.94  2.29  2.41  8.72  3.71
```

We believe these follow an exponential distribution:

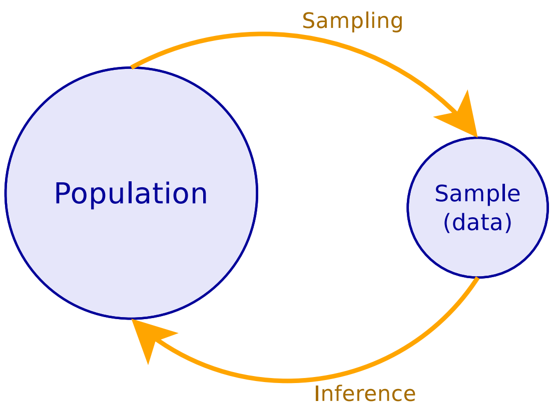
$$X_i \sim \mathrm{Exp}(\lambda)$$

What can we say about λ?

Can we approximate it from the data?

Yes! We can do it using a statistic. This is called *estimation*.

Statistics: the big picture



We want to start learning how to do inference. First, we need a good understanding of the 'sampling' part.

Distributions of statistics

Consider sampling from $X \sim \text{Exp}(\lambda = 1/5)$.

Convenient simplification: set $\theta = 1/\lambda$. This makes $\mathbb{E}(X) = \theta$ and $\text{var}(X) = \theta^2$.

Note: There are two common parameterisations,

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \in [0, \infty)$$
$$f_X(x) = \frac{1}{\theta} e^{-\frac{1}{\theta} x}, \quad x \in [0, \infty)$$

λ is called the *rate parameter* (relates to a Poisson process)

Be clear about which is being used!

Take a large number of samples, each of size $n = 100$:

```
1.   1.84  1.19 11.73  5.64 17.98  0.26  ...
2.   2.67  7.15  5.99  1.03  0.65  3.18  ...
3.  16.99  2.15  2.60  5.40  3.64  2.01  ...
4.   2.21  1.54  4.27  5.29  3.65  0.83  ...
5.  12.24  1.59  2.56  1.38  5.72  0.69  ...
...
```

Then calculate some statistics ($\bar{x}$, $x_{(1)}$, $x_{(n)}$, etc.) for each one:

```
     Min.  Median   Mean    Max.
1.   0.02    4.10   5.17   23.96
2.   0.16    4.48   5.84   39.90
3.   0.17    3.39   4.38   15.61
4.   0.03    3.73   5.43   34.02
5.   0.01    3.12   4.71   19.94
...
```

As we continue this process, we get some information on the distributions of these statistics.

Sampling distribution (definition)

Recall that any statistic $T = \phi(X_1, \ldots, X_n)$ is a random variable.

The *sampling distribution* of a statistic is its probability distribution, given an assumed population distribution and a sampling scheme (e.g. random sampling).

Sometimes we can determine it exactly, but often we might resort to simulation.

In the current example, we know that:

$$X_{(1)} \sim \text{Exp}(100\lambda)$$
$$\sum X_i \sim \text{Gamma}(100, \lambda)$$

How to estimate?

Suppose we want to estimate θ from the data. What should we do?

Reminder:

- Population mean, $\mathbb{E}(X) = \theta = 5$
- Population variance, $\text{var}(X) = \theta^2 = 5^2$
- Population standard deviation, $\text{sd}(X) = \theta = 5$

Can we use the sample mean, $\bar{X}$, as an estimate of θ? *Yes!*

Can we use the sample standard deviation, S, as an estimate of θ? *Yes!*

Will these statistics be good estimates? Which one is better? *Let's see...*

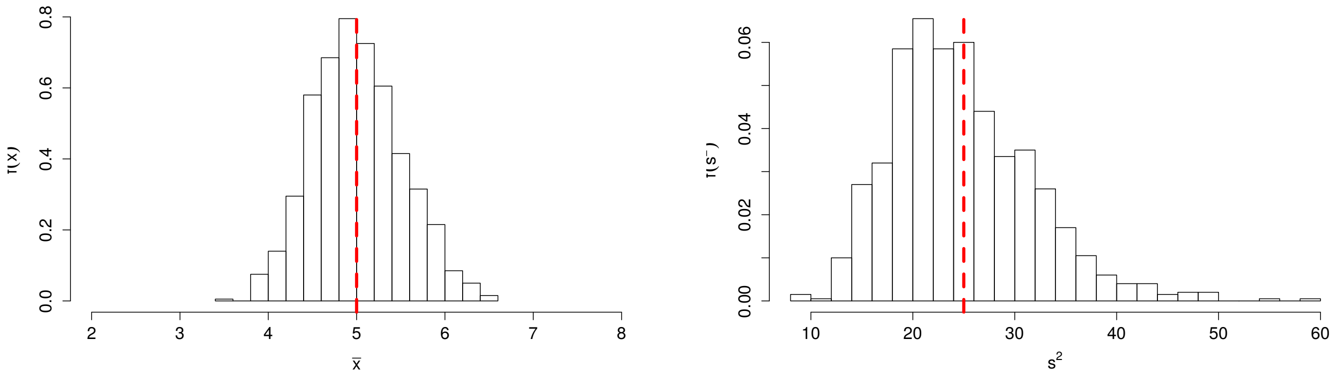
We need to know properties of their sampling distributions, such as their mean and variance.

Note: we are referring to the distribution of the statistic, T, rather than the population distribution from which we draw samples, X.

For example, it is natural to expect that:

- $\mathbb{E}(\bar{X}) \approx \mu$ (sample mean $\approx$ population mean)
- $\mathbb{E}(S^2) \approx \sigma^2$ (sample variance $\approx$ population variance)

Let's see for our example:



Left: distribution of $\bar{X}$. Right: distribution of S^2. Vertical dashed lines: true values, $\mathbb{E}(X) = 5$ and $\text{var}(X) = 5^2$.

- Should we use $\bar{X}$ or S to estimate θ? Which one is the better **estimator**?
- We would like the sample distribution of the estimator to be as close as possible to the true value $\theta = 5$.
- In practice, for any given dataset, we don't know which estimate is the closest, since we don't know the true value.
- We should use the one that is **more likely** to be the closest.
- Simulation: consider 250 samples of size $n = 100$ and compute:

$$\bar{x}_1, \ldots, \bar{x}_{250},$$

$$s_1, \ldots, s_{250}$$

```
> summary(x.bar)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.789   4.663   4.972   5.015   5.365   6.424
> sd(x.bar)
[1] 0.4888185

> summary(s)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.502   4.473   4.916   5.002   5.512   7.456
> sd(s)
[1] 0.7046119
```

From our simulation, $\text{sd}(\bar{X}) \approx 0.49$ and $\text{sd}(S) \approx 0.70$. So, in this case it looks like $\bar{X}$ is superior to S.

2 Estimators

Definitions

- A *parameter* is a quantity that describes the population distribution, e.g. μ and σ^2 for $\text{N}(\mu, \sigma^2)$

- The *parameter space* is the set of all possible values that a parameter might take, e.g. $-\infty < \mu < \infty$ and $0 \leq \sigma < \infty$.
- An *estimator* (or *point estimator*) is a statistic that is used to estimate a parameter. It refers specifically to the random variable version of the statistic, e.g. $T = u(X_1, \dots, X_n)$.
- An *estimate* (or *point estimate*) is the observed value of the estimator for a given dataset. In other words, it is a realisation of the estimator, e.g. $t = u(x_1, \dots, x_n)$, where $x_1, \dots, x_n$ is the observed sample (data).
- *'Hat' notation*: If T is an estimator for θ, then we usually refer to it by $\hat{\theta}$ for convenience.

Examples

We will now go through a few important examples:

- Sample mean
- Sample variance
- Sample proportion

In each case, we assume a sample of iid rvs, $X_1, \dots, X_n$, with mean μ and variance σ^2.

Sample mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Properties:

- $\mathbb{E}(\bar{X}) = \mu$
- $\operatorname{var}(\bar{X}) = \frac{\sigma^2}{n}$

Also, the Central Limit Theorem implies that usually:

$$\bar{X} \approx \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Often used to estimate the population mean, $\hat{\mu} = \bar{X}$.

Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Properties:

- $\mathbb{E}(S^2) = \sigma^2$
- $\operatorname{var}(S^2) =$ (a messy formula)

Often used to estimate the population variance, $\hat{\sigma}^2 = S^2$.

Sample proportion

For a discrete random variable, we might be interested in how often a particular value appears. Counting this gives the *sample frequency*:

$$\operatorname{freq}(a) = \sum_{i=1}^{n} I(X_i = a)$$

Let the *population proportion* be $p = \Pr(X = a)$. Then we have:

$$\operatorname{freq}(a) \sim \operatorname{Bi}(n, p)$$

Divide by the sample size to get the *sample proportion.* This is often used as an estimator for the population proportion:

$$\hat{p} = \frac{\text{freq}(a)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = a)$$

For large n, we can approximate this with a normal distribution:

$$\hat{p} \approx \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$$

Note:

- The sample pmf and the sample proportion are the same, both of them estimate the probability of a given event or set of events.
- The pmf is usually used when the interest is in many different events/values, and is written as a function, e.g. $\hat{p}(a)$.
- The proportion is usually used when only a single event is of interest (getting heads for a coin flip, a certain candidate winning an election, etc.).

Examples for a normal distribution

If the sample is drawn from a normal distribution, $X_i \sim \mathrm{N}(\mu, \sigma^2)$, we can derive **exact** distributions for these statistics.

Sample mean:

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Sample variance:

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

$$\mathbb{E}(S^2) = \sigma^2, \qquad \mathrm{var}(S^2) = \frac{2\sigma^4}{n-1}$$

χ^2_k is the chi-squared distribution with k degrees of freedom. (more details in Module 3)

Bias

Consider an estimator $\hat{\theta}$ of θ.

- If $\mathbb{E}(\hat{\theta}) = \theta$, the estimator is said to be *unbiased*
- The *bias* of the estimator is, $\mathbb{E}(\hat{\theta}) - \theta$

Examples:

- The sample variance is unbiased for the population variance, $\mathbb{E}(S^2) = \sigma^2$. (problem 5 in week 3 tutorial)
- What if we divide by n instead of $n-1$ in the denominator?

Transformations and biasedness

$\mathbb{E}(\frac{n-1}{n}S^2) = \frac{n-1}{n}\sigma^2 < \sigma^2$

$\Rightarrow$ biased!

In general, if $\hat{\theta}$ is unbiased for θ, then it will usually be the case that $g(\hat{\theta})$ is biased for $g(\theta)$.

Unbiasedness is not preserved under transformations.

Challenge problem

Is the sample standard deviation, $S = \sqrt{S^2}$, biased for the population standard deviation, σ?

Choosing between estimators

- Evaluate and compare the sampling distributions of the estimators.
- Generally, prefer estimators that have **smaller bias** and **smaller variance** (and it can vary depending on the aim of your problem).
- Sometimes, we only know asymptotic properties of estimators (will see examples later).

Note: this approach to estimation is referred to as *frequentist* or *classical* inference. The same is true for most of the techniques we will cover. We will also learn about an alternative approach, called *Bayesian* inference, later in the semester.

Challenge problem (uniform distribution)

Take a random sample of size n from the uniform distribution with pdf:

$$f(x) = 1 \quad (\theta - \frac{1}{2} < x < \theta + \frac{1}{2})$$

Can you think of some estimators for θ? What is their bias and variance?

Challenge problem (boundary problem)

Take a random sample of size n from the shifted exponential distribution, with pdf:

$$f(x) = e^{-(x-\theta)} \quad (x > \theta)$$

Equivalently:

$$X_i \sim \theta + \mathrm{Exp}(1)$$

Can you think of some estimators for θ? What is their bias and variance?

Coming up with (good) estimators?

How can we do this for any given problem?

We will cover two general methods:

- Method of moments
- Maximum likelihood

3 Method of moments

Method of moments (MM)

- Idea:
  - Make the population distribution resemble the empirical (data) distribution...
  - ...by *equating theoretical moments with sample moments*
  - Do this until you have enough equations, and then solve them
- Example: if $E(\bar{X}) = \theta$, then the method of moments estimator of θ is $\bar{X}$.
- General procedure (for r parameters):
  1. $X_1, \ldots, X_n$ i.i.d. $f(x \mid \theta_1, \ldots, \theta_r)$.
  2. kth moment is $E(X^k)$
  3. kth sample moment is $M_k = \frac{1}{n} \sum X_i^k$
  4. Set $E(X^k) = M_k$, for $k = 1, \ldots r$ and solve for $(\theta_1, \ldots, \theta_r)$.
- Alternative: Can use the variance instead of the second moment (sometimes more convenient).

Remarks

- An intuitive approach to estimation
- Can work in situations where other approaches are too difficult
- Usually biased
- Usually not optimal (but may suffice)
- Note: some authors use a 'bar' ($\bar{\theta}$) or a 'tilde' ($\widetilde{\theta}$) to denote MM estimators rather than a 'hat' ($\hat{\theta}$). This helps to distinguish different estimators when comparing them to each other.

Example: Geometric distribution

- Sampling from: $X \sim \text{Geom}(p)$
- The first moment:

$$E(X) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = \frac{1}{p}$$

- The MM estimator is obtained by solving

$$\bar{X} = \frac{1}{p}$$

which gives

$$\widetilde{p} = \frac{1}{\bar{X}}$$

Example: Normal distribution

- Sampling from: $X \sim \text{N}(\mu, \sigma^2)$
- Population moments: $\mathbb{E}(X) = \mu$ and $\mathbb{E}(X^2) = \sigma^2 + \mu^2$
- Sample moments: $M_1 = \bar{X}$ and $M_2 = \frac{1}{n}\sum X_i^2$
- Equating them:

$$\bar{X} = \mu \quad \textsf{and} \quad \frac{1}{n}\sum X_i^2 = \sigma^2 + \mu^2$$

Solving these gives:

$$\widetilde{\mu} = \bar{X} \quad \textsf{and} \quad \widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Note:

- This not the usual sample variance!
- $\widetilde{\sigma}^2 = \frac{n-1}{n} S^2$
- This one is biased, $\mathbb{E}(\widetilde{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$.

Example: Gamma distribution

- Sampling from: $X \sim \text{Gamma}(\alpha, \theta)$
- The pdf is:

$$f(x \mid \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} \exp\left(\frac{-x}{\theta}\right)$$

- Population moments: $\mathbb{E}(X) = \alpha\theta$ and $\text{var}(X) = \alpha\theta^2$
- Sample moments: $M = \bar{X}$ and $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$

Equating them:

$$\bar{X} = \alpha\theta \quad \text{and} \quad S^2 = \alpha\theta^2$$

Solving these gives:

$$\widetilde{\theta} = \frac{S^2}{\bar{X}} \quad \text{and} \quad \widetilde{\alpha} = \frac{\bar{X}^2}{S^2}$$

Note:

- This is an example of using S^2 instead of M_2

4 Maximum likelihood estimation

Method of maximum likelihood (ML)

- Idea: find the *'most likely' explanation for the data*
- More concretely: find parameter values that *maximise the probability of the data*

Example: Bernoulli distribution

- Sampling from: $X \sim \mathrm{Be}(p)$
- Data are 0's and 1's
- Then pmf is

$$f(x \mid p) = p^x(1-p)^{1-x}, \quad x = 0, 1, \quad 0 \le p \le 1$$

- Observe values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ (iid)
- The probability of the data (the random sample) is

$$\begin{aligned}\Pr(X_1 = x_1, \ldots, X_n = x_n \mid p) = \prod_{i=1}^{n} f(x_i \mid p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i}\end{aligned}$$

- Regard the sample $x_1, \ldots, x_n$ as known (since we have observed it) and regard the probability of the data as a function of p.
- When written this way, this is called the *likelihood* of p:

$$\begin{aligned}L(p) &= L(p \mid x_1, \ldots, x_n) \\ &= \Pr(X_1 = x_1, \ldots, X_n = x_n \mid p) \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i}\end{aligned}$$

- Want to find the value of p that maximizes this likelihood.
- It often helps to find the value of θ that maximizes the **log** of the likelihood rather than the likelihood
- This is called the *log-likelihood*

$$\ln L(p) = \ln p^{\sum x_i} + \ln(1-p)^{n-\sum x_i}$$

- The final answer (the maximising value of p) is the same, since the log of non-negative numbers is a one-to-one function whose inverse is the exponential, so any value θ that maximises the log-likelihood also maximises the likelihood.
- Putting $x = \sum_{i=1}^{n} x_i$ so that x is the number of 1's in the sample,

$$\ln L(p) = x \ln p + (n-x)\ln(1-p)$$

- Find the maximum of this log-likelihood with respect to p by differentiating and equating to zero,

$$\frac{\partial \ln L(p)}{\partial p} = x\frac{1}{p} + (n-x)\frac{-1}{1-p} = 0$$

- This gives $p = x/n$
- Therefore, the *maximum likelihood estimator* is $\hat{p} = X/n = \bar{X}$

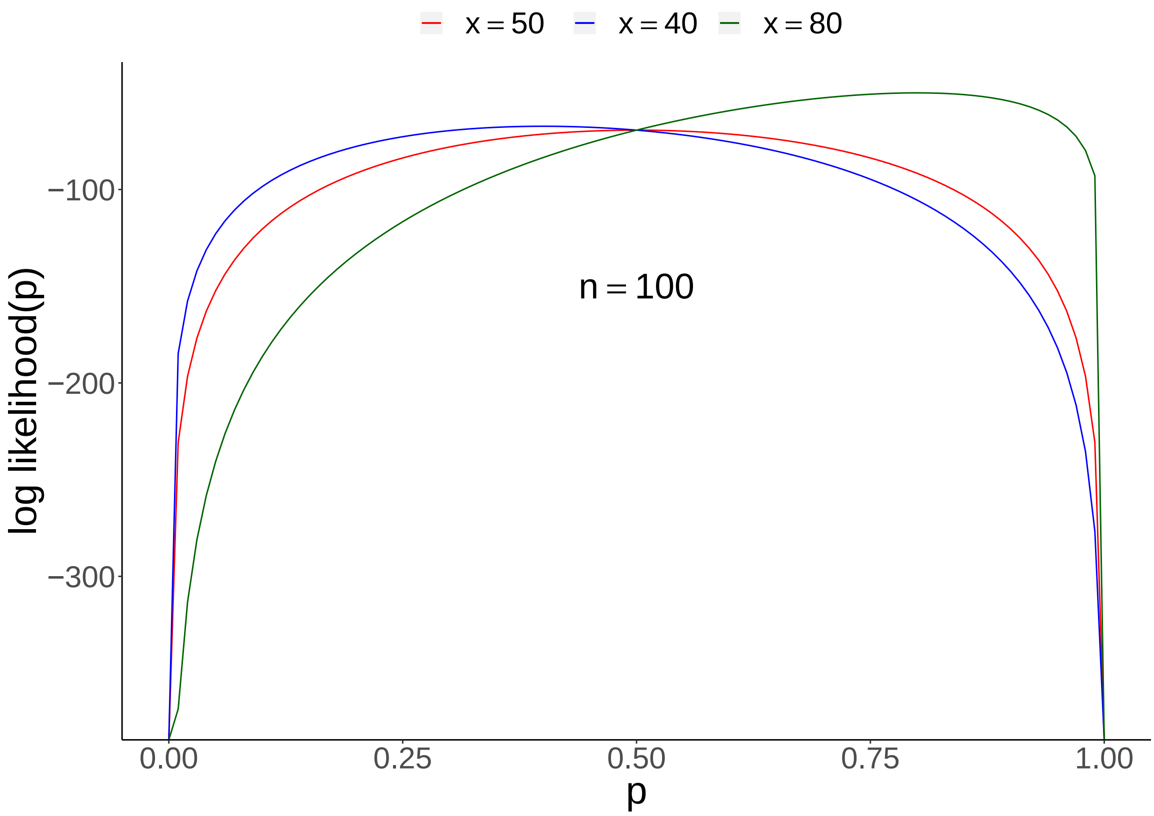


Figure 1: Log-likelihoods for Bernoulli trials with parameter p

Maximum likelihood: general procedure

- Random sample (iid): $X_1, \ldots, X_n$
- Likelihood function with m parameters $\theta_1, \ldots, \theta_m$ and data $x_1, \ldots, x_n$ is:

$$L(\theta_1, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \ldots, \theta_m)$$

- If X is discrete, for f use the pmf
- If X is continuous, for f use the pdf
- The *maximum likelihood estimates* (MLEs) or the *maximum likelihood estimators* (MLEs) $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are values that maximize $L(\theta_1, \ldots, \theta_m)$.
- Note: same abbreviation and notation for both the *estimators* (random variable) and the *estimates* (realised values).
- Often (but not always) useful to take logs and then differentiate and equate derivatives to zero to find MLE's.
- Sometimes this is too hard, but we can maximise numerically. No closed-form expression in this case.

Example: Exponential distribution

Sampling (iid) from: $X \sim \text{Exp}(\lambda)$

$$f(x \mid \lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x > 0, \quad 0 < \lambda < \infty$$

$$L(\lambda) = \frac{1}{\lambda^n} \exp\left(\frac{-\sum_{i=1}^{n} x_i}{\lambda} \right)$$

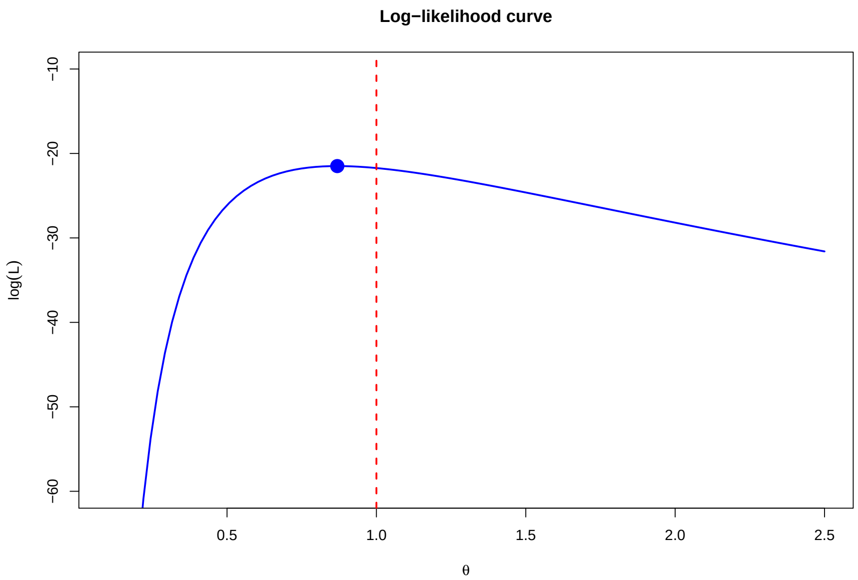
$$\ln L(\lambda) = -n \ln(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} x_i$$

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = -\frac{n}{\lambda} + \frac{\sum x_i}{\lambda^2} = 0$$

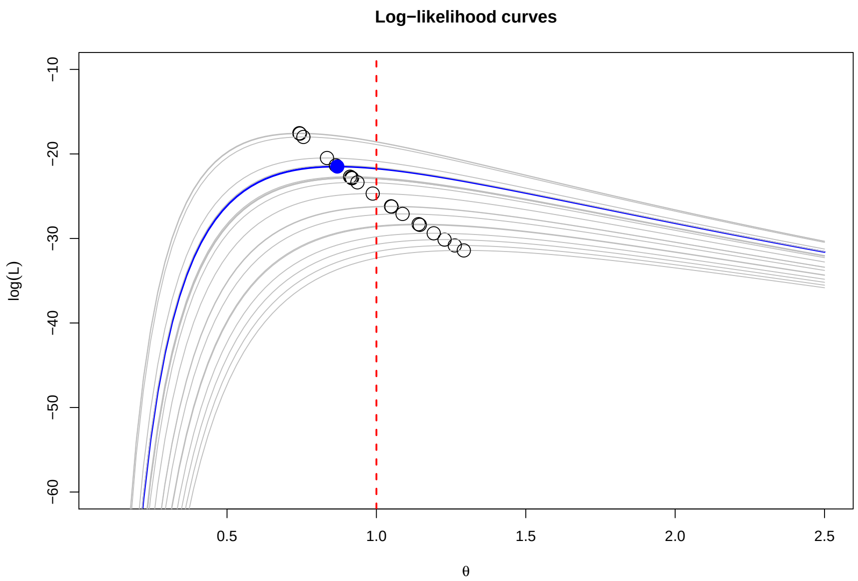
This gives: $\hat{\lambda} = \bar{X}$

Example: Exponential distribution (simulated)

```
> x <- rexp(25)  # simulate 25 observations from Exp(1)
> x
 [1] 0.009669867 3.842141708 0.394267770 0.098725403
 [5] 0.386704987 0.024086824 0.274132718 0.872771164
 [9] 0.950139285 0.022927997 1.538592014 0.837613769
[13] 0.634363088 0.494441270 1.789416017 0.503498224
[17] 0.000482703 1.617899321 0.336797648 0.312564298
[21] 0.702562098 0.265119483 3.825238461 0.238687987
[25] 1.752657238
> mean(x)  # maximum likelihood estimate
[1] 0.8690201
```



What if we repeat the sampling process several times?



Example: Geometric distribution

Sampling (iid) from: $X \sim \text{Geom}(p)$

$$L(p) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum x_i - n}, \quad 0 \le p \le 1$$

$$\frac{\partial \ln L(p)}{\partial p} = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - n}{1-p} = 0$$

This gives: $\hat{p} = 1/\bar{X}$

Example: Normal distribution

Sampling (iid) from: $X \sim \text{N}(\theta_1, \theta_2)$

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{(x_i - \theta_1)^2}{2\theta_2}\right]$$

$$\ln L(\theta_1, \theta_2) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

Take partial derivatives with respect to θ_1 and θ_2.

$$\begin{aligned} \frac{\partial \ln L(\theta_1, \theta_2)}{\partial \theta_1} &= \frac{1}{\theta_2}\sum_{i=1}^{n}(x_i - \theta_1) \\ \frac{\partial \ln L(\theta_1, \theta_2)}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)^2 \end{aligned}$$

Set both of these to zero and solve. This gives: $\theta_1 = \bar{x}$ and $\hat{\theta}_2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The maximum likelihood estimators are therefore:

$$\widehat{\theta}_1 = \bar{X}, \quad \widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

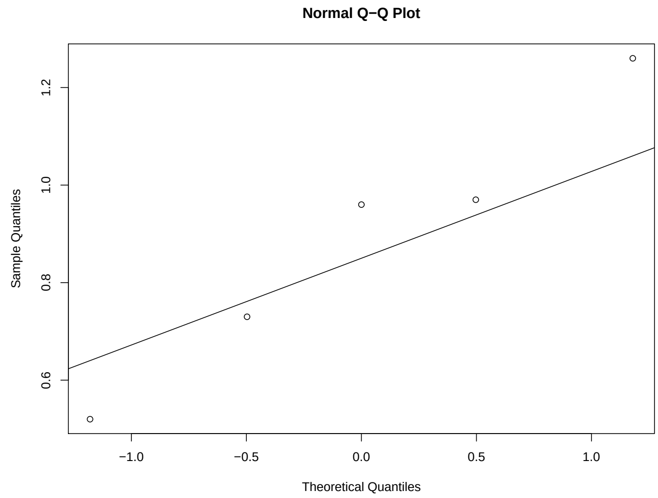
Note: $\widehat{\theta}_2$ is biased.

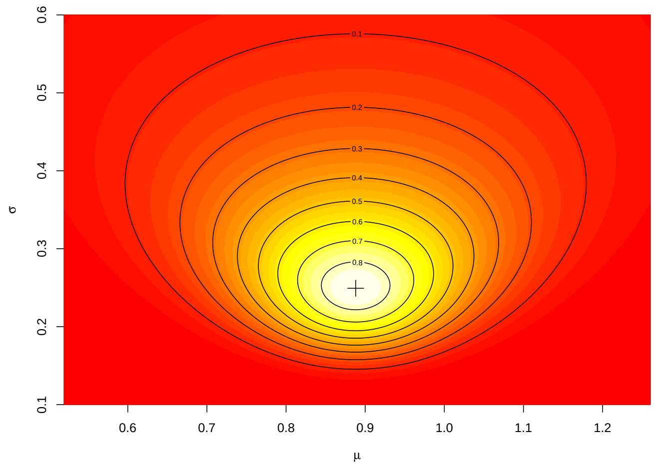
Stress and cancer: VEGFC

```
> x <- c(0.97, 0.52, 0.73, 0.96, 1.26)
> n <- length(x)
> mean(x)  # MLE for population mean
[1] 0.888

> sd(x) * sqrt((n - 1) / n)  # MLE for the pop. st. dev.
[1] 0.2492709

> qqnorm(x)  # Draw a QQ plot
> qqline(x)  # Fit line to QQ plot
```





Challenge problem (boundary problem)

Take a random sample of size n from the shifted exponential distribution, with pdf:

$$f(x \mid \theta) = e^{-(x-\theta)} \quad (x > \theta)$$

Equivalently:

$$X_i \sim \theta + \text{Exp}(1)$$

Derive the MLE for θ. Is it biased? Can you create an unbiased estimator from it?

Invariance property

Suppose we know $\hat{\theta}$ but are actually interestd in $\phi = g(\theta)$ rather than θ itself. Can we estimate ϕ?

Yes! It is simply $\hat{\phi} = g(\hat{\theta})$.

This is known as the *invariance property of the MLE*. In other words, transformations don't affect the value of the MLE.

Consequence: MLEs are usually biased since expectations are **not** invariant under transformations.

Is the MLE a good estimator?

Some useful results:

- Asymptotically unbiased
- Asymptotically optimal variance ('*efficient*')
- Asymptotically normally distributed

The proofs of these rely on the CLT. More details of the mathematical theory will be covered towards the end of the semester.