# Maximum Likelihood Part One[1]

STA312 Spring 2019

[1] See last slide for copyright information.

# Background Reading

- STA256/260 text on maximum likelihood.
- STA258 text or lecture slides on confidence intervals and hypothesis tests.
- Chapter One from *Data analysis with SAS*.

# Overview

1. Background
2. Maximum Likelihood
3. Confidence Intervals and Tests
4. Continuous Models
5. Large Samples

# Statistical Estimation and Inference

- You want to learn from data.
- Adopt a probability model for the data.
- Often, pretend your data are sampled randomly from some population.
- In rare cases, this may even be true.
- What you wish you knew is represented by one or more *unknown parameters*.
- Estimate the parameters, or draw conclusions about the parameters.
- Interpret the results in terms of the data.

# Examples of probability models

Also called *Statistical models*

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value $\mu$ and variance $\sigma^2$.
  The parameters $\mu$ and $\sigma^2$ are unknown.

- For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where
  $\beta_0$ and $\beta_1$ are unknown constants.
  $x_1, \ldots x_n$ are known, observable constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  $\sigma^2$ is an unknown constant.

# Meaning of the regression model

For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where

$\beta_0$ and $\beta_1$ are unknown constants.

$x_1, \ldots x_n$ are known, observable constants.

$\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.

The parameters $\beta_0, \beta_1, \sigma^2$ are unknown constants.

The regression model means

- The predictor $x$ has a rough linear connection to the outcome $y$.
- If $\beta_1 > 0$, low $x$ goes with low $y$ and high $x$ goes with high $y$.
- If $\beta_1 < 0$, low $x$ goes with high $y$ and high $x$ goes with low $y$.
- If $\beta_1 = 0$, then $x$ and $y$ are independent.

# Maximum Likelihood

## Thank you Mr. Fisher

- Denote the unknown parameter by $\theta$.
- How should we estimate $\theta$ based on the sample data?
- Choose the value of $\theta$ that yields the greatest probability of getting the observed data.

# Likelihood

Assuming independent observations (a “random sample”)

$$L(\theta) = \prod_{i=1}^{n} p(y_i|\theta) \text{ or } \prod_{i=1}^{n} f(y_i|\theta)$$

- The likelihood is the probability of obtaining the observed data – expressed as a function of the parameter.
- If the assumed distribution of the data is discrete, this statement is exactly correct.
- If the assumed distribution of the data is continuous, the likelihood is roughly proportional to the probability of observing the data.
- This is a standard calculus problem in maximizing a function.
- It is usually more convenient to maximize the natural log of the likelihood.
- The answer is the same because $\log(x)$ is an increasing function.
- The greater the likelihood, the greater the log likelihood.

# Mechanics

## Really basic math

I have noticed that a major obstacle for many students when doing maximum likelihood calculations is a set of basic mathematical operations they actually know. But the mechanics are rusty, or the notation used in statistics is troublesome. So, with sincere apologies to those who don't need this, here are some basic rules.

# The distributive law

$a(b+c) = ab + ac$. You may see this in a form like

$$\theta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \theta x_i$$

# Power of a product is the product of powers

$(ab)^c = a^c\, b^c$. You may see this in a form like

$$\left(\prod_{i=1}^{n} x_i\right)^{\alpha} = \prod_{i=1}^{n} x_i^{\alpha}$$

# Multiplication is addition of exponents

$a^b a^c = a^{b+c}$. You may see this in a form like

$$\prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n \exp(-\theta \sum_{i=1}^{n} x_i)$$

# Powering is multiplication of exponents

$(a^b)^c = a^{bc}$. You may see this in a form like

$$(e^{\mu t + \frac{1}{2}\sigma^2 t^2})^n = e^{n\mu t + \frac{1}{2}n\sigma^2 t^2}$$

# Log of a product is sum of logs

log means *natural* log, base $e$, possibly denoted ln on your calculator

$\log(ab) = \log(a) + \log(b)$. You may see this in a form like

$$\log \prod_{i=1}^{n} x_i = \sum_{i=1}^{n} \log x_i$$

# Log of a power is the exponent times the log

$\log(a^b) = b\,\log(a)$. You may see this in a form like

$$\log(\theta^n) = n \log \theta$$

# The log is the inverse of the exponential function

$\log(e^a) = a$. You may see this in a form like

$$\log\left(\theta^n \exp(-\theta \sum_{i=1}^{n} x_i)\right) = n \log\theta - \theta \sum_{i=1}^{n} x_i$$

# Example: Coffee taste test

A fast food chain is considering a change in the blend of coffee beans they use to make their coffee. To determine whether their customers prefer the new blend, the company plans to select a random sample of $n = 100$ coffee-drinking customers and ask them to taste coffee made with the new blend and with the old blend, in cups marked "$A$" and "$B$." Half the time the new blend will be in cup $A$, and half the time it will be in cup $B$. Management wants to know if there is a difference in preference for the two blends.

# Statistical model

Letting $\theta$ denote the probability that a consumer will choose the new blend, treat the data $Y_1, \dots, Y_n$ as a random sample from a Bernoulli distribution. That is, independently for $i = 1, \dots, n$,

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

for $y_i = 0$ or $y_i = 1$, and zero otherwise.

# Find the MLE of $\theta$

## Show your work

Denoting the likelihood by $L(\theta)$ and the log likelihood by $\ell(\theta) = \log L(\theta)$, maximize the log likelihood.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \frac{\partial}{\partial \theta} \log \left( \prod_{i=1}^{n} p(y_i|\theta) \right) \\
&= \frac{\partial}{\partial \theta} \log \left( \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \right) \\
&= \frac{\partial}{\partial \theta} \log \left( \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} \right) \\
&= \frac{\partial}{\partial \theta} \left( (\sum_{i=1}^{n} y_i) \log \theta + (n - \sum_{i=1}^{n} y_i) \log(1-\theta) \right) \\
&= \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1-\theta}
\end{aligned}
$$

# Setting the derivative to zero and solving

- $\theta = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$
- Second derivative test: $\frac{\partial^2 \log \ell}{\partial \theta^2} = -n \left( \frac{1-\overline{y}}{(1-\theta)^2} + \frac{\overline{y}}{\theta^2} \right) < 0$
- Concave down, maximum, and the MLE is the sample proportion: $\widehat{\theta} = \overline{y} = p$

# Numerical estimate

Suppose 60 of the 100 consumers prefer the new blend. Give a point estimate the parameter $\theta$. Your answer is a number.

```
> p = 60/100; p
[1] 0.6
```

# Minus log likelihood measures lack of model fit

- $-\ell(\theta) = -\log \prod_{i=1}^{n} p(x_i|\theta) = \sum_{i=1}^{n} -\log p(x_i|\theta)$
- The best fit for observation $x_i$ is if $p(x_i|\theta) = P(X_i = x_i|\theta) = 1$.
- Then the log is zero.
- If $p(x_i|\theta) < 1$, then $\log p(x_i|\theta)$ is negative and $-\log p(x_i|\theta)$ is positive.
- The lower the probability (bad fit), the greater $-\log p(x_i|\theta)$ becomes.
- So maximum likelihood is minimizing the total (or average) badness of fit.
- In machine learning, the minus log likelihood would be called a *loss function*.
- And estimating $\theta$ by minimizing the loss function would be called *learning* about $\theta$.

# Large-sample Normality

## Leading to confidence intervals and tests

- For the taste test example, have MLE $\widehat{\theta} = \overline{y}$, the sample mean.
- The Central Limit Theorem says that if $y_1, \ldots, y_n$ are independent random variables from a distribution with expected value $\mu$ and variance $\sigma^2$, then
- The distribution of $\overline{y}_n$ is approximately normal for large samples.
- Regardless of sample size, $E(\overline{y}_n) = \mu$ and $Var(\overline{y}_n) = \frac{\sigma^2}{n}$.
- Here, the data are Bernoulli, with $\mu = \theta$ and $\sigma^2 = \theta(1-\theta)$.
- Write
$$\widehat{\theta}_n \stackrel{.}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{n}\right).$$
- Vocabulary: $\widehat{\theta}_n$ is "asymptotically normal," with asymptotic mean $\theta$ and asymptotic variance $\theta(1-\theta)/n$.

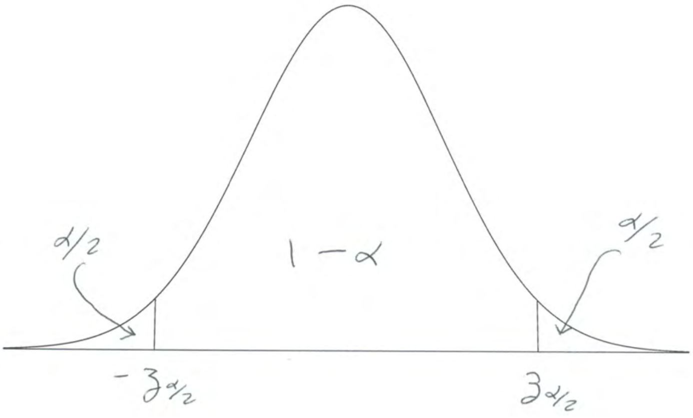
# Large-sample Normality

## Still for the taste test example

- $\widehat{\theta}_n \stackrel{.}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$.
- This means $Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \stackrel{.}{\sim} N(0,1)$.
- Also, $Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}} \stackrel{.}{\sim} N(0,1)$.
- In general, substitute the MLE for the parameter in the formula for the variance, and the Central limit Theorem still holds.
- Substituting the sample variance $s^2$ for $\sigma^2$ also works.

$$s^2 = \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n-1}$$

# Getting the picture



$Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}} \dot{\sim} N(0,1)$ means $P\{-z_{\alpha/2} < Z_n < z_{\alpha/2} \approx 1-\alpha\}$.

# Confidence interval using $Z_n = \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}} \stackrel{.}{\sim} N(0,1)$

$$
\begin{aligned}
1-\alpha &\approx P\{-z_{\alpha/2} < Z_n < z_{\alpha/2}\} \\
&= P\{-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}} < z_{\alpha/2}\} \\
&= P\{-z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} < \widehat{\theta}_n - \theta < z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \\
&= P\{-\widehat{\theta}_n - z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} < -\theta < -\widehat{\theta}_n + z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \\
&= P\{\widehat{\theta}_n + z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} > \theta > \widehat{\theta}_n - z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} \\
&= P\{\widehat{\theta}_n - z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}
\end{aligned}
$$

# Numerical confidence interval for the taste test

Using $1-\alpha \approx P\{\widehat{\theta}_n - z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\sqrt{\frac{\widehat{\theta}_n(1-\widehat{\theta}_n)}{n}}$

```
> thetahat = 60/100; n = 100
> zcrit = qnorm(0.975); zcrit

[1] 1.959964
```

```
> se = sqrt(thetahat*(1-thetahat)/n)
> c(thetahat - zcrit*se, thetahat + zcrit*se)

[1] 0.5039818 0.6960182
```

Confidence interval is $\widehat{\theta} \pm z_{\alpha/2} \times$ standard error.

# Tests of statistical hypotheses

- Model: $y \sim F_\theta$
- $y$ is the data vector, and $\mathcal{Y}$ is the sample space: $y \in \mathcal{Y}$
- $\theta$ is the parameter, and $\Theta$ is the parameter space: $\theta \in \Theta$
- Null hypothesis is $H_0 : \theta \in \Theta_0$ v.s. $H_1 : \theta \in \Theta \cap \Theta_0^c$.
- Meaning of the *null* hypothesis is that *nothing* interesting is happening.
- $\mathcal{C} \subset \mathcal{Y}$ is the *critical region*. Reject $H_0$ in favour of $H_A$ when $y \in \mathcal{C}$.
- Significance level $\alpha$ (*size* of the test) is the maximum probability of rejecting $H_0$ when $H_0$ is true. Conventionally, $\alpha = 0.05$.
- $p$-value is the smallest value of $\alpha$ for which $H_0$ can be rejected.
- Small $p$-values are interpreted as providing stronger evidence against the null hypothesis.

# Carry out a test to determine which brand of coffee is preferred

Recall the model is $y_1, \ldots, y_n \overset{i.i.d.}{\sim} B(1, \theta)$

Start by stating the null hypothesis.

- $H_0 : \theta = 0.50$
- $H_1 : \theta \neq 0.50$
- Could you make a case for a one-sided test?
- $\alpha = 0.05$ as usual.
- Reject $H_0$ if $p < 0.05$.

# Several valid test statistics for $H_0 : \theta = \theta_0$ are available

## Based on $\overline{y} \stackrel{.}{\sim} N(\theta, \frac{\theta(1-\theta)}{n})$

Two of them are

$$Z_1 = \frac{\sqrt{n}(\overline{y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$$

and

$$Z_2 = \frac{\sqrt{n}(\overline{y} - \theta_0)}{\sqrt{\overline{y}(1-\overline{y})}}$$

What is the critical value? Your answer is a number.

```
> alpha = 0.05
> qnorm(1-alpha/2)
[1] 1.959964
```

# Calculate the test statistic and the $p$-value for each test

Suppose 60 out of 100 preferred the new blend

$$Z_1 = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$$

```
> theta0 = .5; ybar = .6; n = 100
> Z1 = sqrt(n)*(ybar-theta0)/sqrt(theta0*(1-theta0)); Z1
[1] 2
> pval1 = 2 * (1-pnorm(Z1)); pval1
[1] 0.04550026
```

$$Z_2 = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\overline{Y}(1-\overline{Y})}}$$

```
> Z2 = sqrt(n)*(ybar-theta0)/sqrt(ybar*(1-ybar)); Z2
[1] 2.041241
> pval2 = 2 * (1-pnorm(Z2)); pval2
[1] 0.04122683
```

# Conclusions

- Do you reject $H_0$? *Yes, just barely.*
- Isn't the $\alpha = 0.05$ significance level pretty arbitrary? *Yes, but if people insist on a Yes or No answer, this is what you give them.*
- What do you conclude, in symbols? $\theta \neq 0.50$. *Specifically,* $\theta > 0.50$.
- What do you conclude, in plain language? Your answer is a statement about coffee. *More consumers prefer the new blend of coffee beans.*
- Can you really draw directional conclusions when all you did was reject a non-directional null hypothesis? *Yes.*

# A technical issue

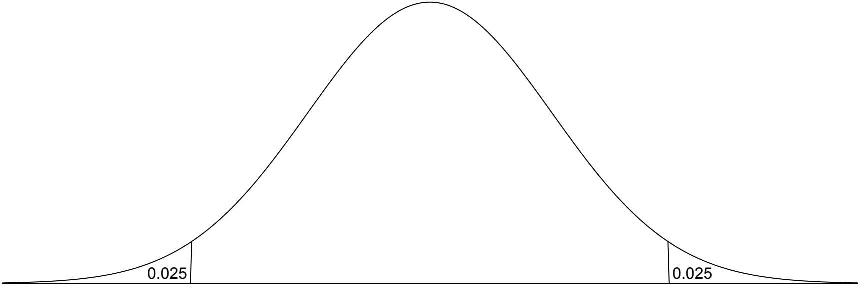
- In this class we will mostly avoid one-tailed tests.
- Why? Ask what would happen if the results were strong and in the opposite direction to what was predicted (dental example).
- But when $H_0$ is rejected, we still draw directional conclusions.
- For example, if $x$ is income and $y$ is credit card debt, we test $H_0 : \beta_1 = 0$ with a two-sided $t$-test.
- Say $p = 0.0021$ and $\widehat{\beta}_1 = 1.27$. We say "Consumers with higher incomes tend to have more credit card debt."
- Is this justified? We'd better hope so, or all we can say is "There is a connection between income and average credit card debt."
- Then they ask: "What's the connection? Do people with lower income have more debt?"
- And you have to say "Sorry, I don't know."
- It's a good way to get fired, or at least look silly.

# The technical resolution

Decompose the two-sided test into a set of two one-sided tests with significance level $\alpha/2$, equivalent to the two-sided test.

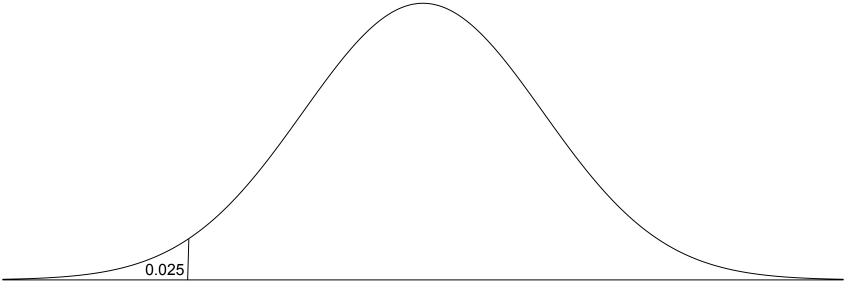
# Two-sided test

$$H_0 : \theta = \tfrac{1}{2} \text{ versus } H_1 : \theta \neq \tfrac{1}{2},\ \alpha = 0.05$$



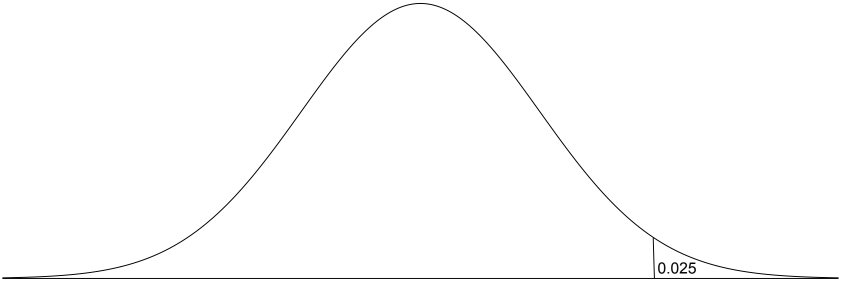
# Left-sided test

$$H_0 : \theta \geq \frac{1}{2} \text{ versus } H_1 : \theta < \frac{1}{2},\ \alpha = 0.025$$



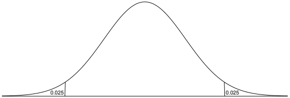
# Right-sided test

$$H_0 : \theta \leq \tfrac{1}{2} \text{ versus } H_1 : \theta > \tfrac{1}{2},\ \alpha = 0.025$$

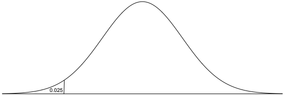


# Decomposing the 2-sided test into two 1-sided tests

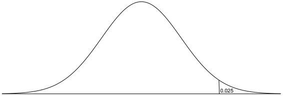
$H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta \neq \frac{1}{2}$, $\alpha = 0.05$



$H_0 : \theta \geq \frac{1}{2}$ vs. $H_1 : \theta < \frac{1}{2}$, $\alpha = 0.025$



$H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$, $\alpha = 0.025$



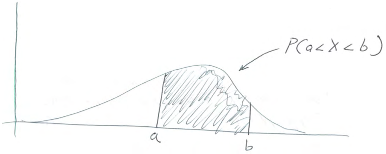
- Clearly, the 2-sided test rejects $H_0$ if and only if exactly *one* of the 1-sided tests reject $H_0$.
- Carry out *both* of the one-sided tests.
- Draw a directional conclusion if $H_0$ is rejected.

# That was a review of confidence intervals and tests

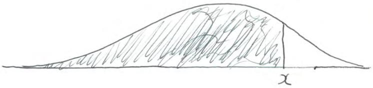
Getting back to maximum likelihood,

# Continuous Random Variable $X$

- Probability is area under a curve.



- The curve is called the *probability density function*.
- It is denoted by $f(x)$ or $f_x(x)$.
- $P(X \le x) = F(x)$ or $F_x(x)$ is the *cumulative distribution function*.

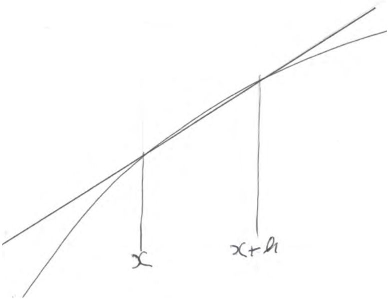


- $\frac{d}{dx}F(x) = f(x)$
- And $F(x) = \int_{-\infty}^{x} f(t)\,dt$

# $f(x) = \frac{d}{dx}F(x)$ is not a probability

Recall $g'(x) = \lim_{h\to 0} \frac{g(x+h)-g(x)}{h}$

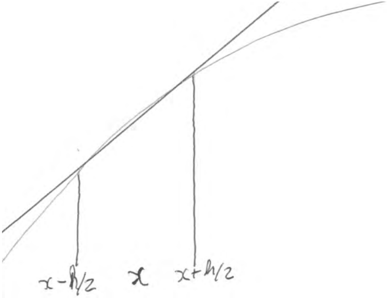
$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h}$$



# Another way to write $f(x)$

Instead of $\lim_{h\to 0} \frac{F(x+h)-F(x)}{h}$

$$f(x) = \lim_{h\to 0} \frac{F(x+\frac{h}{2}) - F(x-\frac{h}{2})}{h}$$



Limiting slope is the same if it exists.

# Interpretation

$$f(x) = \lim_{h \to 0} \frac{F(x + \frac{h}{2}) - F(x - \frac{h}{2})}{h}$$

- $F(x + \frac{h}{2}) - F(x - \frac{h}{2}) = P(x - \frac{h}{2} < X < x + \frac{h}{2})$
- So $f(x)$ is roughly proportional to the probability that $X$ is in a tiny interval surrounding $x$.

# Example: Exponential data

- The lifetime of an electronic component has an exponential distribution with parameter $\lambda > 0$.
- That is, $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x > 0$, and zero for $x \leq 0$.
- Let $X_1, \ldots X_n$ be a random sample of lifetimes.
- What is the likelihood function? Simplify.

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

- Note that $x_1, \ldots, x_n$ are the observed data values.
- The likelihood is roughly proportional to the probability of obtaining a set of data values in a tiny neighbourhood of the observed sample data.

# Find the MLE

## Differentiate the log likelihood

$$\begin{aligned}
\frac{d}{d\lambda}\ell(\lambda) &= \frac{d}{d\lambda}\log L(\lambda) \\
&= \frac{d}{d\lambda}\log\left(\lambda^n e^{-\lambda\sum_{i=1}^n x_i}\right) \\
&= \frac{d}{d\lambda}\left(n\log\lambda - \lambda\sum_{i=1}^n x_i\right) \\
&= \frac{n}{\lambda} - \sum_{i=1}^n x_i \overset{set}{=} 0 \\
\Rightarrow & \quad \lambda = \frac{n}{\sum_{i=1}^n x_i}
\end{aligned}$$

So $\widehat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = 1/\overline{x}$.

# Large-sample normality of the MLE

- For the coffee taste test (Bernoulli) example, the MLE $\widehat{\theta}$ was approximately normal because (in that example) $\widehat{\theta} = \overline{y}$, and the Central Limt Theorem says $\overline{y}$ is approximately normal for large samples.
- But the result holds more generally.
- Under some technical conditions that are satisfied in this class, the distribution of the maximum likelihood estimate is approximately normal for large samples.
- The distribution of *vectors* of parameters is approximately multivariate normal.
- Thank you, Mr. Wald.

# A Central Limit Theorem for the MLE

## Based indirectly on the usual Central Limit Theorem

$$\widehat{\theta}_n \stackrel{.}{\sim} N(\theta, \frac{1}{n\,I(\theta)})$$

Where $I(\theta)$ is the *Fisher Information* in one observation.

$$I(\theta) = E\frac{\partial^2}{\partial\theta^2} - \log f(X|\theta) = -E\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)$$

Here's the idea.

- You are finding the MLE by *minimizing* the *minus* log likelihood function.
- And doing the second derivative test to see if it's really a minimum.
- But the likelihood is a random quantity, because the $X_i$ values are random variables.
- So take the expected value.

# Fisher Information in the whole sample

$I(\theta) = -E\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)$ is the information in one observation.

$$\begin{aligned} -E\frac{\partial^2}{\partial\theta^2}\log L(\theta) &= -E\frac{\partial^2}{\partial\theta^2}\log\prod_{i=1}^{n} f(X_i|\theta) \\ &= -E\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log f(X_i|\theta) \\ &= \sum_{i=1}^{n} -E\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta) \\ &= n\,I(\theta) \end{aligned}$$

# Variance and curvature of the log likelihood

- Fisher observed that some likelihood functions are almost flat at the MLE, while others have a lot of curvature (big second derivative).
- Likelihoods with more curvature contain more information about the location of the parameter.
- The Fisher information in the whole sample (that's $nI(\theta)$) is the expected curvature of the minus log likelihood, at the true parameter value.
- Fisher's great insight was that the curvature is deeply related to the variance of the MLE.
- The more the curvature, the smaller the variance.
- The asymptotic variance of the MLE is $v_n = \frac{1}{nI(\theta)}$.
- For many examples it's exactly the variance.
- Fisher discovered this. Wald proved asymptotic normality under general conditions.

# Estimating the asymptotic variance $v_n = \frac{1}{nI(\theta)}$

- For tests and confidence intervals, we need to *estimate* the asymptotic variance of the MLE.
- There are (at least) two good ways.
- The first is to use $\frac{1}{nI(\widehat{\theta})}$.
- The other estimate is based on the Fisher information in the whole sample: $nI(\theta) = -E\frac{\partial^2}{\partial\theta^2}\log L(\theta) = -E\frac{\partial^2}{\partial\theta^2}\ell(\theta)$.
- Instead of calculating the expected value and then substituting $\theta = \widehat{\theta}$, just substitute $\theta = \widehat{\theta}$ in the first place.
- The result is sometimes called the *observed* Fisher information:

$$\widehat{nI(\theta)} = -\frac{\partial^2}{\partial\theta^2}\log L(\theta)\Big|_{\theta=\widehat{\theta}} = -\ell''\left(\widehat{\theta}\right)$$

- Often, the two estimates are identical. They are always close for large samples.

# Observed Fisher Information: $-\ell''(\widehat{\theta})$

## The second derivative of the minus log likelihood, evaluated at the MLE

We now have a convenient recipe for the standard error (estimated standard deviation) of the MLE.

- Differentiate the log likelihood function and set to zero; solve for the MLE.
- Carry out the second derivative test to make sure it's a maximum.
- That is, differentiate again and substitute $\theta = \widehat{\theta}$.
- Multiply by minus one and invert it (one over). That's the estimated variance of the MLE.
- Take the square root, and you have the standard error.

$$S_{\widehat{\theta}} = \frac{1}{\sqrt{-\ell''(\widehat{\theta})}}$$

# What do you need to be able to do?

Given a model, and a set of numerical data,

- Derive a formula for $\widehat{\theta}$.
- Calculate a numerical point estimate of $\theta$ from the sample data. The answer is a number.
- Give a formula for the estimated variance of $\widehat{\theta}$. Use $\widehat{v}_n = 1/-\ell''(\widehat{\theta})$.
- Calculate a 95% confidence interval for $\theta$. Use $1.96 \pm z_{\alpha/2} \times$ standard error. The answer is a pair of numbers.
- Test $H_0 : \theta = \theta_0$. Use
$$Z_n = \frac{\widehat{\theta} - \theta_0}{\sqrt{\widehat{v}_n}}.$$
- We need an example.

# Copyright Information