Elements of statistics (MATH0487-1)

Prof. Dr. Dr. K. Van Steen

University of Liège, Belgium

November 12, 2012

Outline I

Outline II

Outline III

Outline IV

Types of Estimators

- Focus on finding **point estimators** first, i.e. for which the true value of a (function of a) parameter is assumed to be a point.
- Several methods exist to compute point estimators, including the "methods of moments" and "maximum likelihood", but also the "method of least squares" (see Regression Analysis chapter), etc.
- Second, focus on finding **interval estimators**, i.e. acknowledge the utility for some interval about the point estimate together with some measure of accurance that the true value of the parameter lies within the interval.

Types of Estimators

Inference choices:

1. making the inference of estimating the true value of the parameter to be a point,
2. making the inference of estimating that the true value of the parameter is contained in some interval.

Desirable Properties of a Point Estimator

- **Unbiased functions**: $\hat{g}(\mathbf{X})$ is said to be unbiased for a function $g(\theta)$ if $\mathrm{E}[\hat{g}(\mathbf{X})] = g(\theta)$.
- Even if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ will generally not be an unbiased estimator of $g(\theta)$ unless g is linear or affine. We need additional properties ...

Unbiasedness	
Trading off Bias and Variance	MSE MVUE
Efficiency	
Consistency	
Sufficiency	

- Depending on the method used to derive a point estimator, different performances are to be expected

Method 1: Sample Moments as Estimators

- For a random variable X, the rth moment about the origin 0, or the rth moment of its corresponding density function is defined as $\mu'_r = E(X^r)$.
- For a random sample $X_1, X_2, \ldots, X_n$, the rth sample moment about the origin is defined by

$$M_r = \sum_{i=1}^{n} X_i^r / n, r = 1, 2, 3, \ldots$$

  and its observed value is denoted by $m_r = \sum_{i=1}^{n} x_i^r / n$.
- The following property of sample moments holds:

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample of X. Then

$$E(M_r) = \mu'_r, r = 1, 2, 3, \ldots$$

The MME Procedure

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $F(x : \theta_1, \ldots, \theta_k)$. Hence, suppose that there are k parameters to be estimated.
- Let μ'_r, m_r $(r = 1, 2, \ldots, k)$ denote the first k population and sample moments respectively.
- Suppose that each of these population moments are certain *known* functions of the parameters:

$$\begin{aligned}
\mu'_1 &= g_1(\theta_1, \ldots, \theta_k), \\
\mu'_2 &= g_2(\theta_1, \ldots, \theta_k), \\
&\vdots \\
\mu'_k &= g_k(\theta_1, \ldots, \theta_k).
\end{aligned}$$

- Solving simultaneously the set of equations,

$$g_r(\hat{\theta}_1, \ldots, \hat{\theta}_k) = m_r, r = 1, 2, \ldots, k$$

gives the required estimates $\hat{\theta}_1, \ldots, \hat{\theta}_k$.

The MME Procedure Applied

- Let μ_r', m_r $(r = 1, 2)$ denote the first k population and sample moments. The population moments are known functions of these population parameters.
- For the normal distribution, we know that $\mu_1' = E(X) = \mu$ and $\sigma^2 = E(X^2) - \mu^2$, so $\mu_2' = E(X^2) = \sigma^2 + \mu^2$.
- The unknown parameters to estimate are μ and σ^2
- Equate:

$$m_1 = \frac{1}{n}\sum x_i = \overline{x} \rightarrow \mu_1' = \mu,$$

$$m_2 = \frac{1}{n}\sum x_i^2 \rightarrow \mu_2' = \sigma^2 + \mu^2.$$

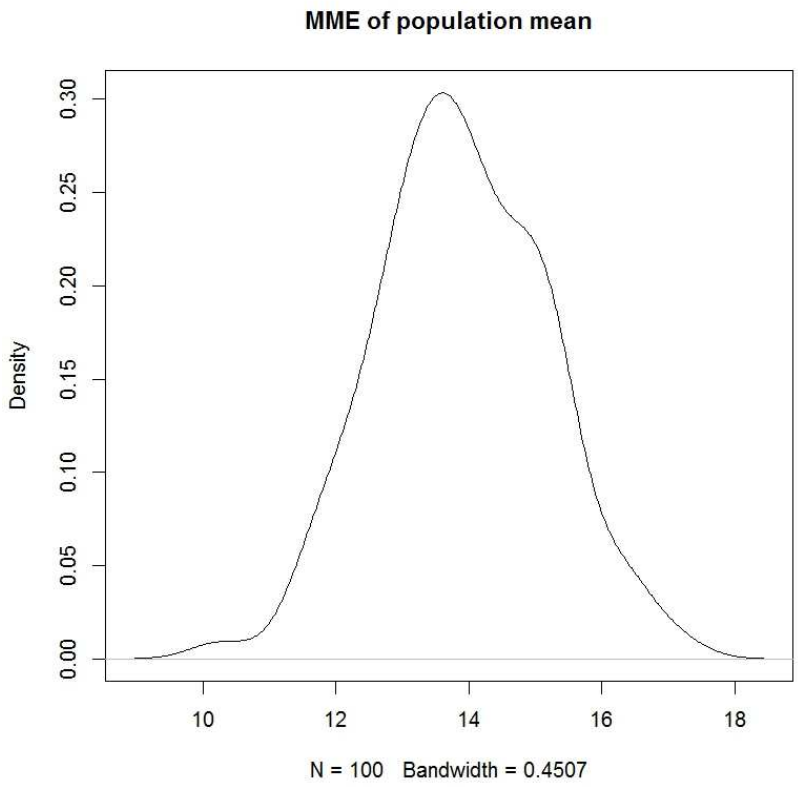
- Solving simultaneously the set of equations, gives

$$\hat{\mu} = \overline{x}, \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2.$$

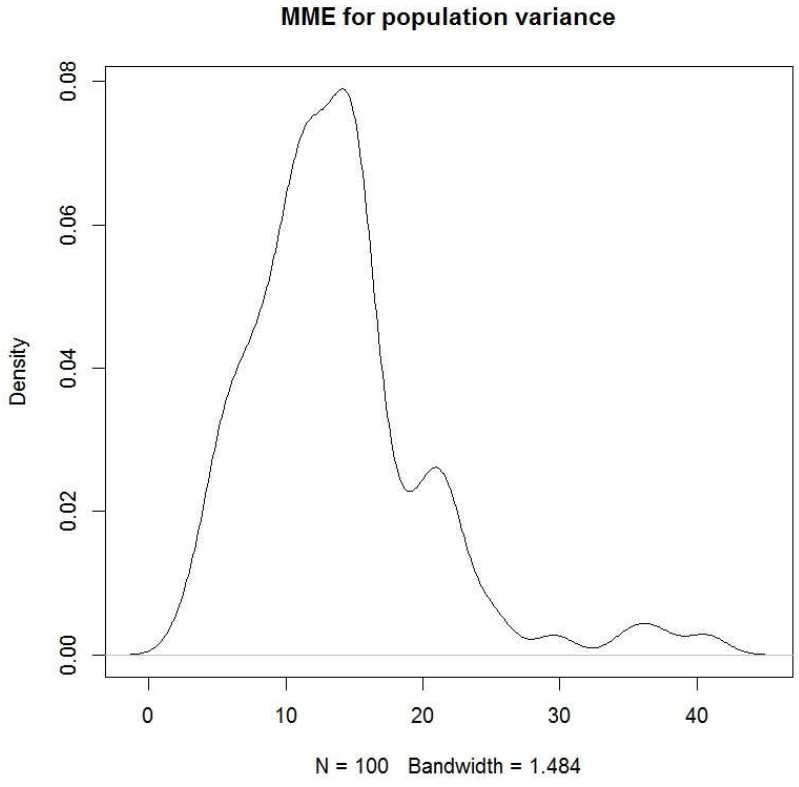
The MME Procedure Applied

```
#_____________NormalMoments.R ___________
set.seed(69)
mu <- 14
sigma <- 4
sampsz <- 10
nsimulations <- 100
mu.estimates <- numeric(nsimulations)
var.estimates <- numeric(nsimulations)
for (i in 1:nsimulations){
rn <- rnorm(mean=mu,sd=sigma,n=sampsz)

## computing MMEs
mu.estimates[i] <- mean(rn)
var.estimates[i] <- mean( (rn -mean(rn))^2 )

} # end of i loop

plot(density(mu.estimates),main="MME of population mean")
plot(density(var.estimates),main="MME for population variance")
```

The MME Procedure Applied



pretty "ok"; MME=$\overline{x}$



pretty "skew"; MME=$\frac{1}{n}\sum x_i^2 - \overline{x}^2$

Mean-Squared Error

Lemma (The MSE variance-bias tradeoff)

The MSE, $E((\hat{\theta} - \theta)^2)$, decomposes as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2.$$

Uniformly minimum-variance unbiased estimator - UMVUE

- Why is the MME method a good starting point? Answer: The MME method almost always produces some asymptotically unbiased estimators, although they may not be the best estimators.
- When no bias-variance trade-off can be made, one approach is to restrict ourselves to the subclass of estimators that are *unbiased* and *minimum variance*
- If an unbiased estimator of $g(\theta)$ has minimum variance among all unbiased estimators of $g(\theta)$ it is called a **minimum variance unbiased estimator** (MVUE).

How to find MVUE when it exists?

Lemma (Fisher information)

The variance of $S(\theta)$ is the expected Fisher information about θ

$$\mathrm{E}(\mathcal{I}(\theta)) = \mathrm{E}\{S(\theta)^2\} \equiv \mathrm{E}\left\{\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right\}$$

- Throughout this course, we will call $S(\theta)^2$ **observed Fisher Information** $\mathcal{I}(\theta)$ and $\mathrm{E}\{S(\theta)^2\}$ (logically) **expected Fisher information** $\mathrm{E}(\mathcal{I}(\theta))$.
- Remember: $\mathrm{E}(\mathcal{I}(\theta)) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right]$

How to find a lower bound for MVUE?

Theorem (Cramér Rao Lower Bound - CRLB)

Let $\hat{\theta}$ be an unbiased estimator of θ. Then

$$\mathrm{Var}(\hat{\theta}) \geq \{ \mathrm{E}(\mathcal{I}(\theta)) \}^{-1}.$$

- Although this bound is typically used to prioritize estimators among the class of unbiased estimators, it is also useful to assess the "quality" of a biased estimator:
- Biased estimators will be considered good, if their variances are lower than the CRLB.
- Remember: Since we are talking about estimators/estimates, we need to compute the expected Fisher information on the sample.... (so do not forget that this will involve the sample size n)

Efficiency

- An unbiased estimator $\hat{\theta}$ is said to be **efficient** if $\mathrm{eff}(\hat{\theta}) = 1$, with

$$\mathrm{eff}(\hat{\theta}) = \frac{\mathrm{CRLB}}{\mathrm{Var}(\hat{\theta})},$$

- The **asymptotic efficiency** of an unbiased estimator $\hat{\theta}$ is the limit of the efficiency as $n \to \infty$.
- For 2 unbiased estimators of θ ($\widehat{\theta_1}$ and $\widehat{\theta_2}$) with respective variances $Var(\widehat{\theta_1})$, $Var(\widehat{\theta_2})$, $\widehat{\theta_1}$ is *more efficient* than $\widehat{\theta_2}$ if

$$Var(\widehat{\theta_1}) < Var(\widehat{\theta_2}).$$

Consistency

- Consistency has to do only with the *limiting behaviour of an estimator* as the sample size increases without limit and does not imply that the observed value of $\widehat{\theta}$ is necessarily close to θ for any specific size of sample n.
- If only a relatively small sample is available, it would seem immaterial whether a consistent estimator is used or not.
- $\widehat{\theta}_n$ is a **consistent** estimator of θ if

$$\lim_{n\to\infty} P(\left|\widehat{\theta}_n - \theta\right| > \epsilon) = 0 \text{ for all } \epsilon > 0.$$

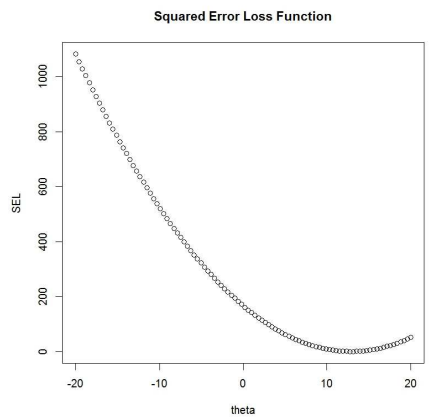
We then say that $\widehat{\theta}_n$ **converges in probability** to θ as $n \to \infty$.

Theorem

If $\lim_{n\to\infty} E(\widehat{\theta}_n) = \theta$ *and* $\lim_{n\to\infty} Var(\widehat{\theta}_n) = 0$, *then* $\widehat{\theta}_n$ *is a consistent estimator of* θ.

The Concepts of Loss and Risk

- Consider estimating θ and let $H(x_1, \ldots, x_n)$ denote an estimate of θ.
- The **loss function** denoted by $l(H(x_1, \ldots, x_n); \theta)$ is defined to be a real-valued function satisfying
  - $l(H(x_1, \ldots, x_n); \theta) \geq 0$ for all possible estimates $H(x_1, \ldots, x_n)$ and all allowable θ
  - $l(H(x_1, \ldots, x_n); \theta) = 0$ for $H(x_1, \ldots, x_n) = \theta$
- Example: $l(H(x_1, \ldots, x_n); \theta) = (\widehat{\theta} - \theta)^2$ (squared error loss)
- The function $l(H(x_1, \ldots, x_n); \theta)$ equals the **loss** incurred if one estimates the true parameter to be $\widehat{\theta}$



The average loss or **risk function** is defined as $\mathcal{R}_l(\theta) = E(l(\widehat{\theta}; \theta))$ and coincides with the MSE for the squared error loss function above.

Sufficiency

- The random sample $X_1, X_2, \ldots, X_n$ drawn from the distribution with $F(x;\theta)$ contains information about the parameter θ.
- To estimate θ, this sample is first condensed to a single random variable by use of a statistic $\theta^* = H(X_1, X_2, \ldots, X_n)$.
- The question now of interest is whether any information about θ has been lost by this condensing process : With a possible choice of $\theta^* = H(X_1, \ldots, X_n) = X_1$, it seems that some of the information in the sample has been lost since the observations $X_2, \ldots, X_n$ have been ignored.
- In many cases, the statistic θ^* *does* contain all the relevant information about the parameter θ that the sample contains, in which case we call it a *sufficient* statistic.

Sufficiency

- Formally, let $X_1, \ldots, X_n$ be a random sample from the density $f(.;\theta)$, where θ may be a vector. A statistic $S(X_1, \ldots, X_n)$ is a **sufficient** statistic if and only if the conditional distribution of $X_1, \ldots, X_n$ given $S = s$ does not depend on θ for any value s of S.
- Without proof, this definition is equivalent to the following one:

  Let $X_1, \ldots, X_n$ be a random sample from the density $f(.;\theta)$, where θ may be a vector. A statistic $S = g(X_1, \ldots, X_n)$ is a **sufficient** statistic if and only if the conditional distribution of T given $S = s$ does not depend on θ, for any statistic $T = t(X_1, \ldots, X_n)$.
- Let $X_1, \ldots, X_n$ be a random sample from the density $f(.;\theta)$, where θ may be a vector. Statistics $S_1, \ldots, S_r$ are said to be **jointly sufficient** if and only if the conditional distribution of $X_1, \ldots, X_n$ given $S_1 = s_1, \ldots, S_r = s_r$ does not depend on θ.

Rao-Blackwell Theorem

- The concept of sufficiency can certainly help us in our search for UMVUEs.
- Loosely speaking, an unbiased estimator which is a function of sufficient statistics has smaller variance than an unbiased estimator which is not based on sufficient statistics.
- Hence, our aim is to look for unbiased estimators that are functions of sufficient statistics ...

Rao-Blackwell Theorem

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from the density $f(.;\theta)$, and let $S_1 = s_1(X_1, X_2, \ldots, X_n), \ldots, S_k = s_k(X_1, X_2, \ldots, X_n)$ be a set of jointly sufficient statistics. Let the statistic $T = t(X_1, X_2, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$. Define T' by $T' = \mathrm{E}(T|S_1, \ldots, S_k)$, then

- *T' is a statistic and a function of the sufficient statistics $S_1, \ldots, S_k$. [Hence, we can write $T' = t'(S_1, \ldots, S_k)$]*
- $\mathrm{E}(T') = \tau(\theta)$. *[Hence, T' is an unbiased estimator of $\tau(\theta)$]*
- $\mathrm{Var}_\theta[T'] \leq \mathrm{Var}_\theta[T]$, *for every θ.*

Sufficiency by example

- Consider three independent binomial trials where $\theta = P(X = 1)$.

Event			Probability	Set
0	0	0	$(1-\theta)^3$	A_0
1	0	0		
0	1	0	$\theta(1-\theta)^2$	A_1
0	0	1		
0	1	1		
1	0	1	$\theta^2(1-\theta)$	A_2
1	1	0		
1	1	1	θ^3	A_3

- $T = t(X) = \sum X_i$ (i.e., the number of "successes") identifies A_i.
- Once we know in which set A_i the sample belongs to, $P(X = x|A_i)$ does not depend on θ : i.e. $P(010|A_1; \theta) = 1/3$.

Sufficiency by example

- Summarizing:

  - You should think of sufficiency in the sense of using all the relevant information in the sample.
  - For example, to say that $\overline{X}$ is sufficient for μ in a particular distribution means that knowledge of the actual observations $x_1, x_2, \ldots, x_n$ gives us no more information about μ than does only knowing the average of the n observations.

How to find a sufficient estimator?

Theorem (Neyman's Factorization Criterion)

A statistic $T = t(X)$ is sufficient for θ if and only if the joint density $f(x;\theta)$ of $X_1, \ldots, X_n$ can be factorized as

$$f(x;\theta) = h(x)k\left\{t(x);\theta\right\}, \; x \in \mathcal{X}, \; \theta \in \Theta.$$

i.e. into a function $h(.)$ which does not depend on θ and a function $k(.)$ which only depends on x through $t(x)$. This is true in general (i.e., can be generalized to find jointly sufficient statistics).

How to find a sufficient estimator?

- Let $X = (X_1, \ldots, X_n)$ be independent and Bernoulli distributed with parameter θ so that

$$f(x;\theta) \;=\; \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \;=\; \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

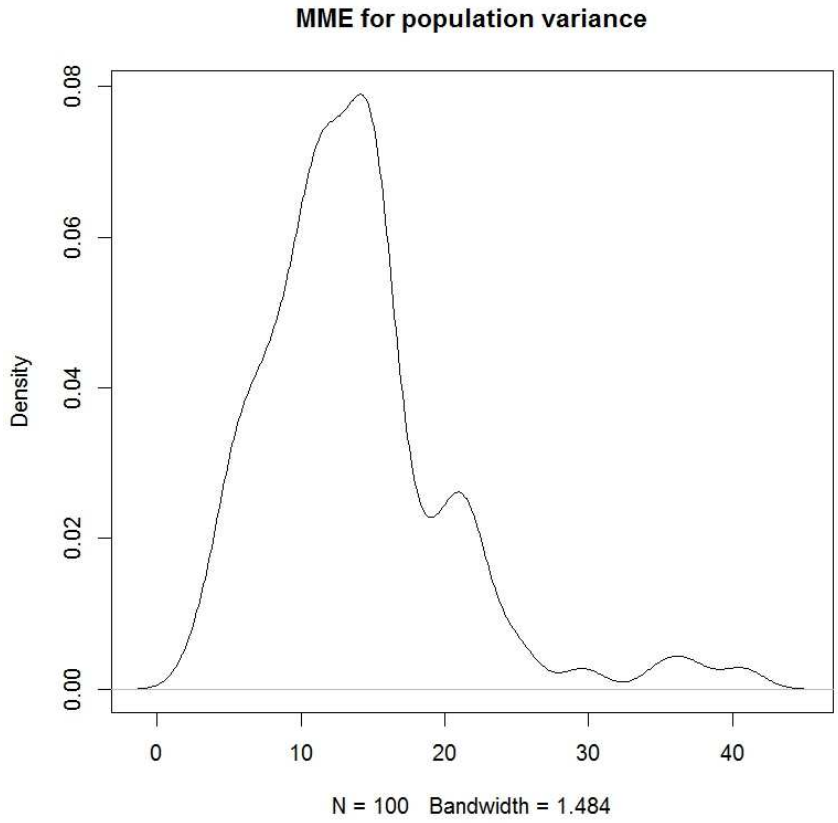
- Take $k\{\sum x_i;\theta\} = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$ and $h(x) = 1$, then $t(x) = \sum_i x_i$ is sufficient.
- In a similar way, it can be shown that $T_1 = t_1(X) = \bar{X}$ is sufficient for μ when $X_i \sim i.i.d.\ \mathrm{N}(\mu, \sigma^2)$. In fact, $\sum X_i$ and $\sum X_i^2$ are jointly sufficient.
- Methods of moments estimators may NOT be functions of sufficient statistics (for instance in the case the uniform distribution family over the interval $[\theta_1, \theta_2]$)
- In contrast, maximum likelihood estimators will ALWAYS depend o the sample through any set of jointly sufficient statistics.

Crude Summary

- *Bias* = expected value of estimator does not necessarily equal parameter
- *Consistency* = estimator approaches parameter as n approaches infinity
- *Efficiency* = smaller variance of parameter implies higher efficiency
- *Sufficient* = utilizes all pertinent information in a sample

Properties of an MME

Estimator	Unbiased	Consistent	Efficient	Sufficient
MME		ν	generally not very efficient	not necessarily
	when adjusted to be unbiased	$\longrightarrow$	often leads to minimum variance estimators	



The Likelihood Function

- Let $x_1, x_2, \ldots, x_n$ be sample observations taken on the random variables $X_1, X_2, \ldots, X_n$. Then the **likelihood of the sample**, $L(\theta|x_1, x_2, \ldots, x_n) = f_X(x|\theta)$, is defined as:
  - the joint probability of $x_1, x_2, \ldots, x_n$ if $X_1, X_2, \ldots, X_n$ are discrete, and
  - the joint probability density function of $X_1, \ldots, X_n$ evaluated at $x_1, x_2, \ldots, x_n$ if the random variables are continuous.

The Likelihood Function

- The **likelihood function** for a set of n identically and independently distributed (i.i.d.) random variables, $X_1, X_2, \ldots, X_n$, can thus be written as:

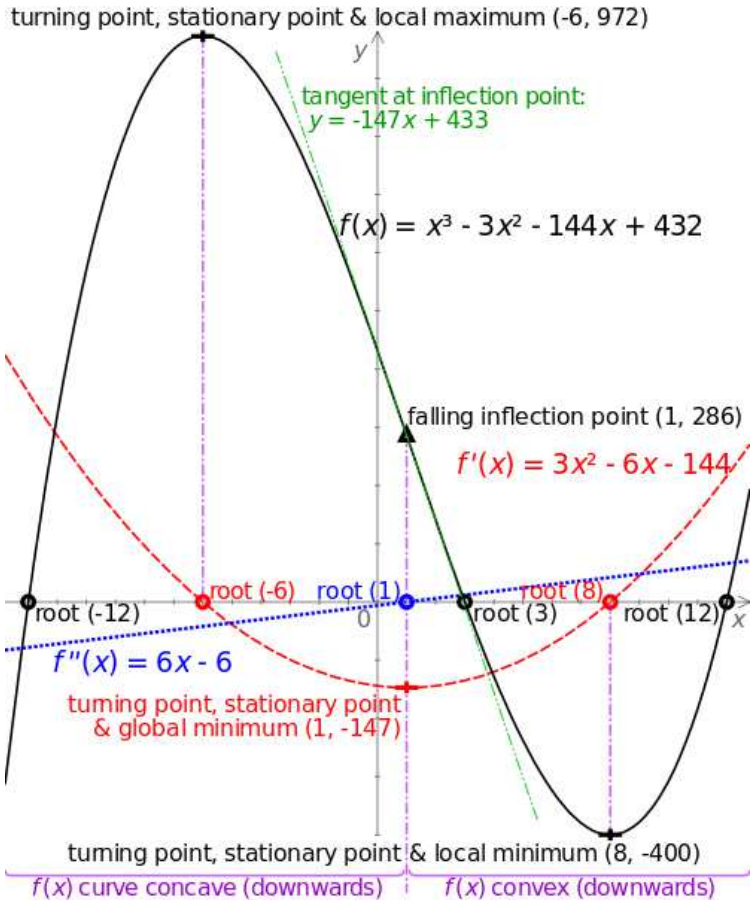
$$Ł(\theta; x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n) & \text{for } X \text{ discrete,} \\ f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta) & \text{for } X \text{ continuous} \end{cases}.$$

- Important: the argument of $f_X(x; \boldsymbol{\theta}) \equiv f_X(x|\boldsymbol{\theta})$ is x, but the argument of $L(\boldsymbol{\theta}|x)$ is $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$ is a vector of m unknown parameters to be estimated.

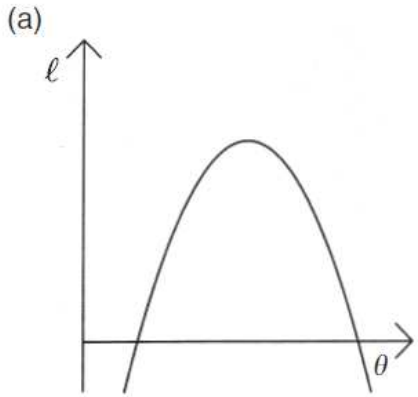
Method 2: Maximum Likelihood Estimation

- The **maximum likelihood estimate** (MLE) $\hat{\theta}$ of θ is the solution to the score equation
$$S(\theta) = 0.$$
- In other words, the maximum likelihood estimate (MLE) of θ is that value of θ which maximizes the likelihood. Or stated otherwise, the MLE of θ is that value of θ, say $\widehat{\theta}$ such that
$$L(\widehat{\theta}; x_1, \ldots, x_n) > L(\theta'; x_1, \ldots, x_n)$$
where θ' is any other value of θ.
- To find MLE:
  - it is helpful to take log
  - one needs calculus (taking derivatives)
  - one should remember to check values at boundaries and second derivatives

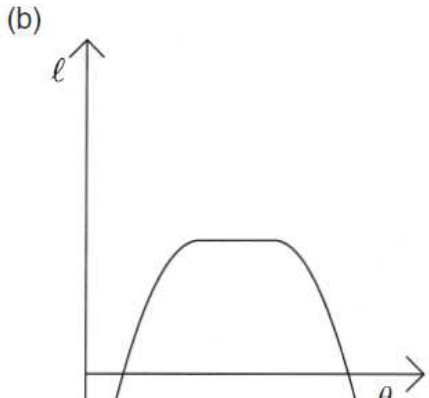
Maximum Likelihood Estimation



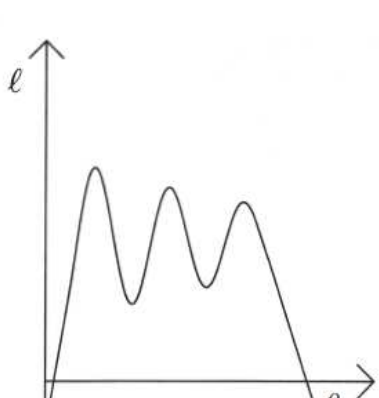
Examples of Likelihood Functions



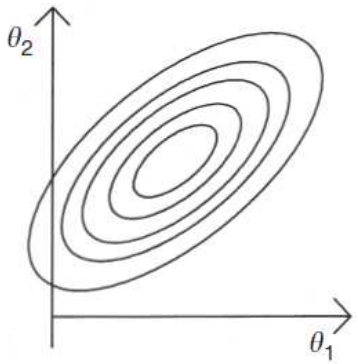
Well-behaved one-dimensional likelihood



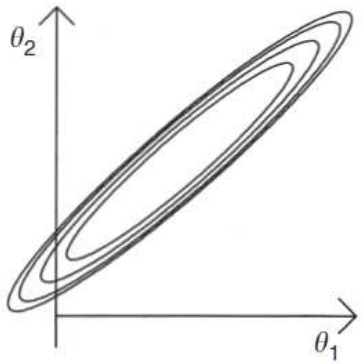
Flat one-dimensional likelihood



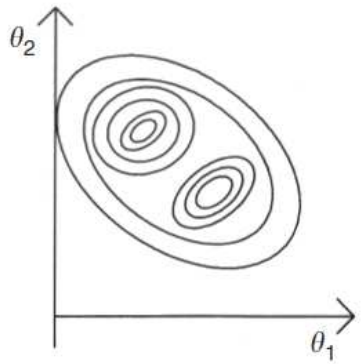
Multimodal one-dimensional likelihood



Well-behaved two-dimensional likelihood



Flat (ridged) two-dimensional likelihood



Multimodal two-dimensional likelihood

Log-Likelihood Function for Gaussian Distribution

- Consider data $X_1, X_2 \ldots, X_n$ distributed as $\mathrm{N}(\mu, \upsilon)$. Then the likelihood function is

$$L(\mu, \upsilon) = \left(\frac{1}{\sqrt{\pi\upsilon}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\upsilon}\right\}$$

- The log-likelihood function is

$$\ell(\mu, \upsilon) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\upsilon) - \frac{1}{2\upsilon}\sum_{i=1}^{n}(x_i - \mu)^2$$

MLEs for Gaussian Distribution

Unknown mean and known variance:

- As υ is known we treat this parameter as a constant when differentiating wrt μ.
- Then

$$S(\mu) = \frac{1}{\upsilon}\sum_{i=1}^{n}(x_i - \mu), \qquad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and

$$\mathrm{E}[\mathcal{I}(\mu)] = n\frac{1}{\upsilon^2}\mathrm{E}[(X-\mu)^2] = \frac{n}{\upsilon} \quad > 0 \ \ \forall\ \mu.$$

- Also, $\mathrm{E}[\hat{\mu}] = n\mu/n = \mu$, and so the MLE of μ is unbiased.
- Finally

$$\mathrm{Var}[\hat{\mu}] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] = \frac{\upsilon}{n} = \left(\mathrm{E}[\mathcal{I}(\theta)]\right)^{-1}.$$

MLEs for Gaussian Distribution

Known mean and unknown variance:

- Differentiating the log-likelihood function for the Gaussian density with mean μ and variance υ wrt υ returns

$$S(\upsilon) = -\frac{n}{2\upsilon} + \frac{1}{2\upsilon^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

and setting $S(\upsilon) = 0$ implies

$$\hat{\upsilon} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

- Differentiating again, and multiplying by -1 yields the information function

$$I(\upsilon) = -\frac{n}{2\upsilon^2} + \frac{1}{\upsilon^3}\sum_{i=1}^{n}(x_i - \mu)^2.$$

[Note: $-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2\right] = \mathrm{E}(\mathcal{I}(\theta))$]

MLEs for Gaussian Distribution

Known mean and unknown variance:

- Now define

$$Z_i = (X_i - \mu)/\sqrt{v},$$

  so that $Z_i \sim \mathrm{N}(0, 1)$.

- It can be shown that

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2; \ \mathrm{E}[\sum Z_i^2] = n, \mathrm{Var}[\sum Z_i^2] = 2n.$$

- Remarks:
  - In fact, the chi-square distribution is defined as a particular function, which we will introduce later.
  - For every density for a random variable U we can compute a **moment generating function**: $m_U(t) = \mathrm{E}_U(e^{tU})$. Differentiating $m_U(t)$ r times wrt t and letting $t \to 0$ gives the rth moment around 0.
  - If two moment generating functions agree on some interval for t, then the corresponding cumulative distributions functions agree.
  - Computing the moment generating function $m_U(t)$ for $U = \sum Z_i^2$ gives the moment generating function of a chi-squared distribution

    ...

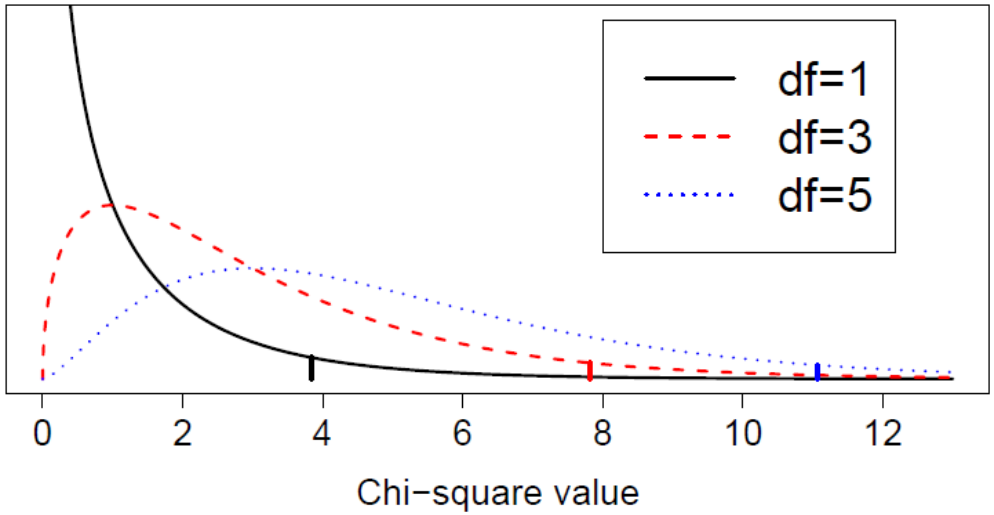
The Chi-square Density Function

- If X is a random variable with density

$$f_X(x) = \frac{1}{\Gamma(k/2)}(\frac{1}{2})^{k/2}x^{k/2-1}e^{-\frac{1}{2x}}, x \geq 0$$

then X is defined to have a **chi-square distribution with k degrees of freedom**. The density given above is called a *chi-square density with k degrees of freedom* (a positive integer).

- Here, $\Gamma(j) = (j-1)!$ (for any positive integer j) and $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt$ (for any complex number z with positive real part)



MLEs for Gaussian Distribution

Known mean and unknown variance:

- Our MLE can therefore be expressed as

$$\hat{v} = (v/n)\sum_{i=1}^{n} Z_i^2,$$

and

$$\mathrm{E}[\hat{v}] = \mathrm{E}\left[\frac{v}{n}\sum_{i=1}^{n} Z_i^2\right] = v,\ \mathrm{Var}[\hat{v}] = \left(\frac{v}{n}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} Z_i^2\right] = \frac{2v^2}{n}.$$

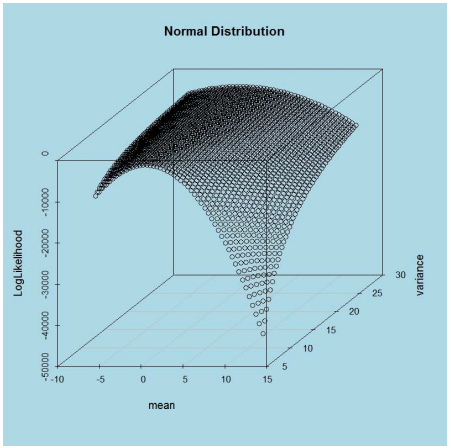
- Finally,

$$\mathrm{E}\left[I(v)\right] = -\frac{n}{2v^2} + \frac{1}{v^3}\sum_{i=1}^{n}\mathrm{E}\left[(x_i-\mu)^2\right] = -\frac{n}{2v^2} + \frac{nv}{v^3} = \frac{n}{2v^2}.$$

Hence the $\mathrm{CRLB} = 2v^2/n$, and so $\hat{v}$ has efficiency 1.

MLEs for Gaussian Distribution

- Our treatment of the two parameters of the Gaussian distribution in the last example was to
  - (i) fix the variance and estimate the mean using maximum likelihood, or
  - (ii) fix the mean and estimate the variance using maximum likelihood.



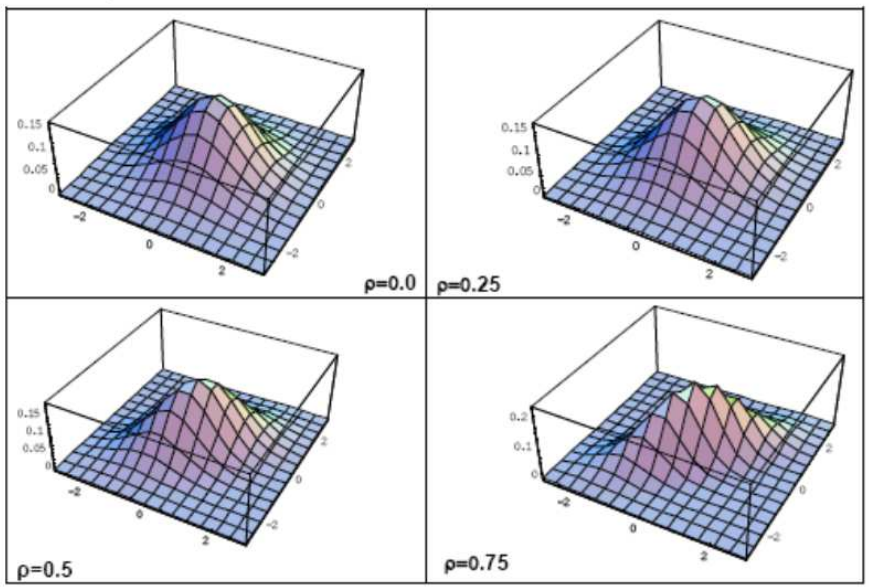
- It is possible to consider simultaneous MLEs of these parameters! - no exam material

Multi-parameter Case getting More Complex

- Under certain conditions, alternatively, the **profile likelihood** may be used for estimation purposes just like any other likelihood.
- Suppose, in the Gaussian example before, the interest is in the mean μ. Then first treat it as fixed and derive the maximum likelihood estimate for σ^2. This MLE is a function of μ.
- Next, substitue σ^2 with this function of μ wherever it occurs in the two-parameter based log-likelihood, giving rise to a function of μ solely, say $L_p(\mu|x_1, \ldots, x_n)$. Obtain the maximum profile likelihood estimate for the mean by derivating $L_p(\mu|x_1, \ldots, x_n)$ for the parameter of interest μ.
- It can be shown that this maximum is exactly the sample mean (hence equal to the overall MLE for the problem that wishes to maximize the log-likelihood surface for μ and σ at the same time).

Multi-parameter Case getting More Complex

Example: The bivariate normal setting



- Bivariate densities add to the complexity: there is also (at least one) parameter "connecting" the two marginal densities corresponding to the two random variables under study

Properties of an MLE

Estimator	Unbiased	Consistent	Efficient	Sufficient
MME		ν	generally not very efficient	not necessarily
	when adjusted to be unbiased		often lead to minimum variance estimators	
MLE	with increasing ss becomes unbiased	ν	with increasing ss becomes MVUE	if MLE exists, then it is suff.

- The maximum likelihood estimate is unique for most "generalized linear models" (see later)
- Having in mind the construction of confidence intervals and genuine hypothesis testing, it is interesting to know that the MLE has a distribution that tends to normality as $n \to \infty$. (See also syllabus: MLEs for means and variances)

Dealing with parameter transformations

- The MLE is invariant under functional transformations: a one-to-one function g evaluated in an MLE estimate for a parameter θ will be an MLE estimate of $g(\theta)$. This is called the **invariance property** of MLEs.
- It seems intuitive that if $\hat{\theta}$ is most likely for θ and our knowledge (data) remains unchanged then $g(\hat{\theta})$ is most likely for $g(\theta)$. Frequentists generally accept the invariance principle without question.
- This is not the case for Bayesians, who assign a probability distribution to a parameter.
- Note that the invariant property is not necessarily true for other (other than MLE) estimators.
- For example, if $\hat{\theta}$ is the MVUE of θ, then $g(\hat{\theta})$ is generally not MVUE for $g(\theta)$.

Interesting functions of MLE estimates

- Now suppose that an experiment consists of measuring random variables $X_1, X_2, \ldots, X_n$ which are i.i.d. with probability distribution depending on a parameter θ.
  - Let $\hat{\theta}$ be the MLE of θ. Define

$$\begin{aligned} W_1 &= \sqrt{\mathrm{E}[I(\theta)]}(\hat{\theta} - \theta) \\ W_2 &= \sqrt{I(\theta)}(\hat{\theta} - \theta) \\ W_3 &= \sqrt{\mathrm{E}[I(\hat{\theta})]}(\hat{\theta} - \theta) \\ W_4 &= \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta). \end{aligned}$$

  - Then, W_1, W_2, W_3, and W_4 are all random variables and, as $n \to \infty$, the probabilistic behaviour of each of W_1, W_2, W_3, and W_4 is well approximated by that of a $N(0, 1)$ random variable.
- As a consequence, for instance:
  - $\mathrm{E}[W_1] \approx 0$, thus $\mathrm{E}[\hat{\theta}] \approx \theta$ [i.e., approx unbiased]
  - $\mathrm{Var}[W_1] \approx 1$, thus $\mathrm{Var}[\hat{\theta}] \approx (\mathrm{E}[I(\theta)])^{-1}$ [i.e., approx efficient]

Sampling Distibutions

- We have seen the following table before. Note that *mean* and *variance* are only a few characteristics, that may or may not completely define a distribution! They do, when the underlying distribution can be assumed to be Gaussian ...

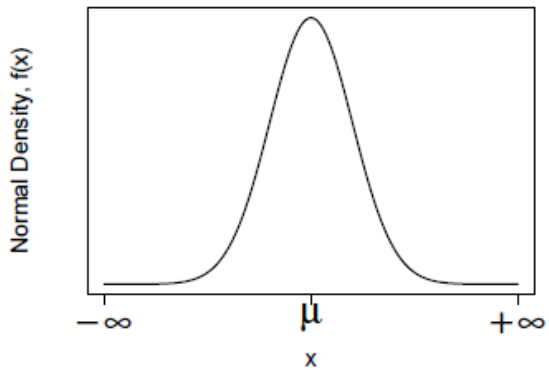
Statistic	Mean	Variance
$\overline{X}$	μ	$\frac{\sigma^2}{n}$
$\overline{X_1} - \overline{X_2}$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

- Sampling distributions allow us to make statements about the *unobserved true population parameter* in relation to the *observed sample statistic* $\longrightarrow$ **statistical inference**

The Normal Distribution

For completion, the normal density function

- takes on values between $-\infty$ and $+\infty$,
- Mean = Median = Mode,
- area under the curve equals 1

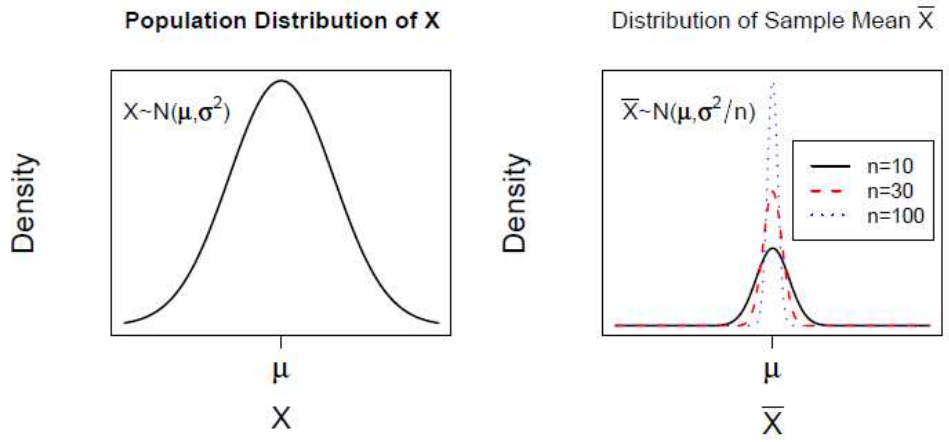


The normal probability density function for $X \sim N(\mu, \sigma^2)$ is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < +\infty$$

Note: $\pi \approx 3.14$ and $e \approx 2.72$ are mathematical constants

Sampling Distribution of the Sample Mean $\overline{X}$



- When sampling from a *normally* distributed population,
  - $\overline{X}$ will be normally distributed
  - The mean of the distribution of $\overline{X}$ is equal to the true mean μ of the population from which the samples were drawn
  - The variance of the distribution is σ^2/n, where σ^2 is the variance of the population and n is the sample size
  - We can write: $\overline{X} \sim N(\mu, \sigma^2/n)$
- When sampling from a population whose distribution is not normal and the sample size is large, use the **Central Limit Theorem**

The Central Limit Theorem

- Given a population of *any* distribution with mean μ and variance σ^2, the sampling distribution of $\overline{X}$, computed from samples of size n from this will be *approximately* $N(\mu, \sigma^2/n)$ when this sample size is large:
- In general, this applies when $n \geq 25$
- The approximation of normality obviously becoms better as n increases

The Standard Normal Distribution

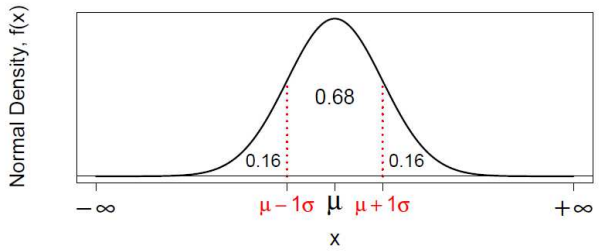
- Definition: a normal distribution $N(\mu, \sigma^2)$ with parameters $\mu = 0$ and $\sigma = 1$
- Its density function is written as

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty$$

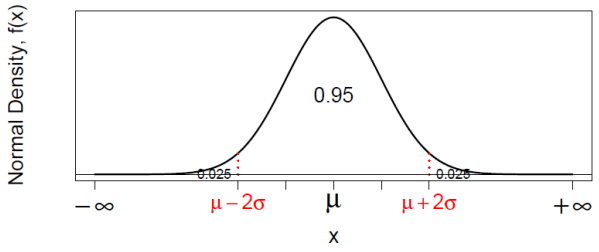
- We typically use the letter Z to denote a **standard normal** random variable: $Z \sim N(0, 1)$
- Important: We can use the standard normal all the time (instead of non-standardized version) because if $X \sim N(\mu, \sigma^2)$ then $\frac{X-\mu}{\sigma} \sim N(0, 1)$
- This process is called "standardizing" a normal random variable

The Standard Normal Distribution: 68-95-99.7 Rules

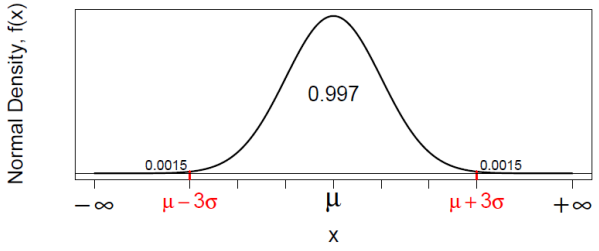
68% of the density is within one standard deviation of the mean



95% of the density is within two standard deviations of the mean



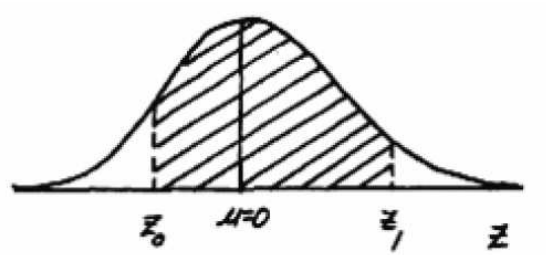
99.7% of the density is within three standard deviations of the mean



Normal Probabilities

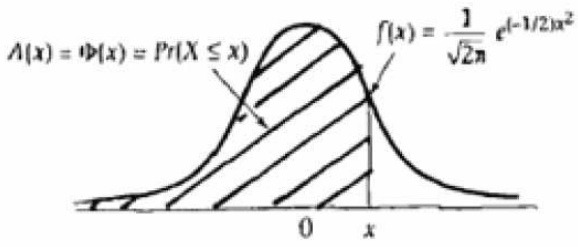
- We are often interested in the probability that z takes on values between z_0 and z_1 (not necessarily symmetric round the mean μ):

$$P(z_0 \leq z \leq z_1) = \int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2} dz$$

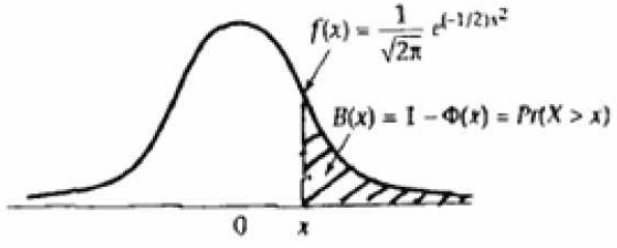


- Do we always have to (re-)compute this integral?

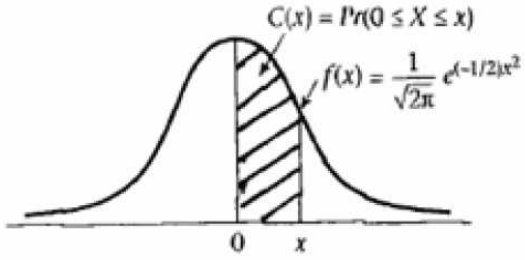
Z Tables



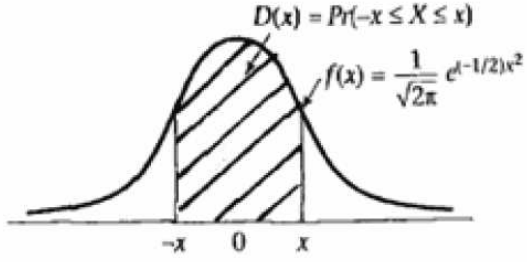
(a)



(b)



(c)



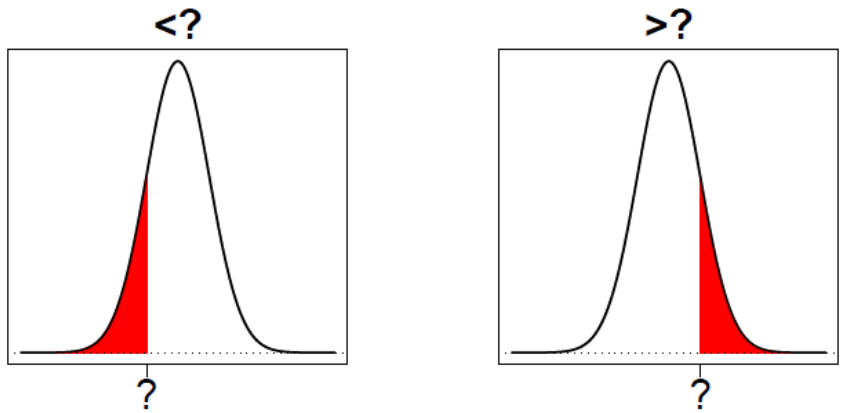
(d)

x	A^a	B^b	C^c	D^d	x	A	B	C	D
0.0	.5000	.5000	.0	.0	0.32	.6255	.3745	.1255	.2510
0.01	.5040	.4960	.0040	.0080	0.33	.6293	.3707	.1293	.2586
0.02	.5080	.4920	.0080	.0160	0.34	.6331	.3669	.1331	.2661
0.03	.5120	.4880	.0120	.0239	0.35	.6368	.3632	.1368	.2737
0.04	.5160	.4840	.0160	.0319	0.36	.6406	.3594	.1406	.2812
0.05	.5199	.4801	.0199	.0399	0.37	.6443	.3557	.1443	.2886
0.06	.5239	.4761	.0239	.0478	0.38	.6480	.3520	.1480	.2961

"Software R" Tables

For **standard normal** random variables $Z \sim N(0,1)$ we'll use

1. `pnorm(?)` to find $P(Z \leq ?)$
2. `pnorm(?, lower.tail=F)` to find $P(Z \geq ?)$



For **any** normal random variable $X \sim N(\mu, \sigma^2)$
(but taking $X \sim N(2, 3^2)$ as an example) we'll use

1. `pnorm(?, mean=2, sd=3)` to find $P(X \leq ?)$
2. `pnorm(?, mean=2, sd=3, lower.tail=F)` to find $P(X \geq ?)$

Normal approximations

- Under certain conditions, the normal distribution can be used to approximate $Binomial(n, p)$ distribution
  - $np > 5$
  - $n(1-p) > 5$
- For instance, $\hat{p} \sim N(P, nP(1-P))$

Statistic	Mean	Variance
$\hat{p}$	P	$\frac{P(1-P)}{n}$
$n\hat{p}$	nP	$nP(1-P)$
$\hat{p}_1 - \hat{p}_2$	$P_1 - P_2$	$\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}$

From Point to Interval Estimators

Point estimation

- $\overline{X}$ is a point estimator of μ
- $\overline{X_1} - \overline{X_2}$ is a point estimator of $\mu_1 - \mu_2$
- $\hat{p}$ is a point estimator of P
- $\hat{p}_1 - \hat{p}_2$ is a point estimator of $P_1 - P_2$

We know the sampling distribution of these statistics, e.g.,

$$\overline{X} \sim N(\mu_{\overline{X}} = \mu, \sigma^2_{\overline{X}} = \sigma^2/n)$$

If σ^2 is not known, we need to estimate it. The natural point estimator for σ^2 is the sample variance s^2.

From Point to Interval Estimators

Interval estimation

- $100(1-\alpha)\%$ confidence interval:

  point estimate $\pm$ (critical value of z or t or ...) $\times$ (standard error)

- Example confidence interval for the population mean (plugging in values):

$$\overline{X} \pm z_{\alpha/2} \times \sigma_{\overline{X}}$$

The $z_{\alpha/2}$ is the value such that under a standard normal curve, the area under the curve that is larger than $z_{\alpha/2}$ is $\alpha/2$ and the area under the curve that is less than $-z_{\alpha/2}$ is $\alpha/2$.

Why do we use z from a standard normal, whereas we know that $\overline{X}$ does not follow a standard normal distribution but has mean $\overline{x}$?

Derivation of Confidence Interval (CI) for the Mean

We get the $100(1-\alpha)\%$ confidence interval for μ by taking:

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= 1-\alpha \\ P(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq z_{\alpha/2}) &= 1-\alpha \\ P(-z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \bar{X}-\mu \leq z_{\alpha/2} \cdot \sigma_{\bar{X}}) &= 1-\alpha \end{aligned}$$

After some algebra:

$$\begin{aligned} P(\bar{X} - z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \sigma_{\bar{X}}) &= 1-\alpha \\ P(L \leq \mu \leq U) &= 1-\alpha \end{aligned}$$

Derivation of Confidence Interval (CI) for the Mean

- So, a $100(1-\alpha)\%$ confidence interval for μ, the population mean, is given by the interval estimate:

$$\overline{x} \pm z_{\alpha/2} \times \sigma_{\overline{x}} \ldots$$

  when the population variance is known!
- However, the population variance is rarely known.

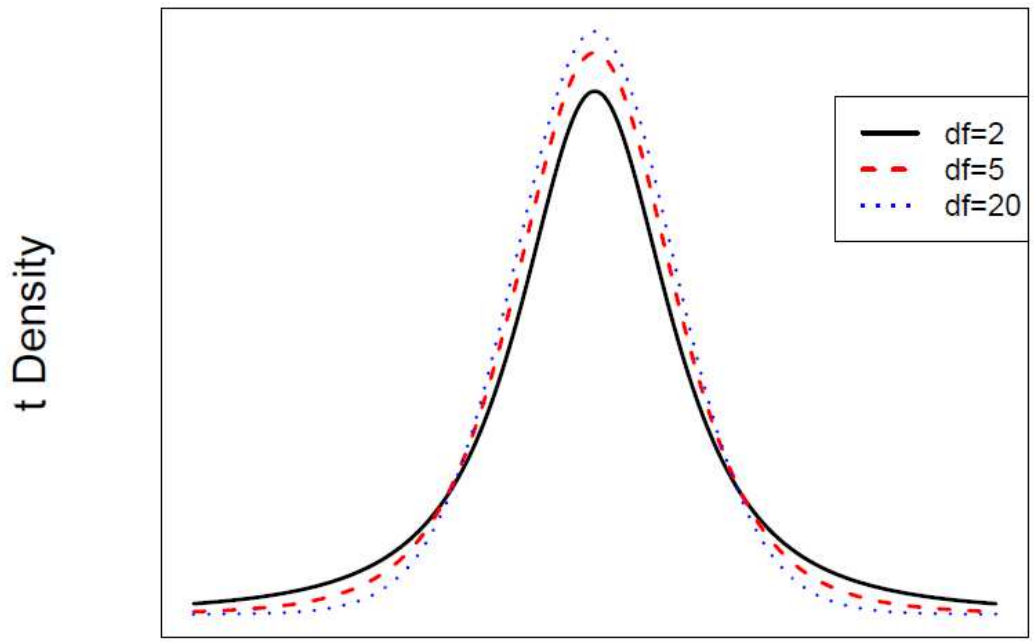
How could we possibly deal with this?

Using the Sample Variance

- Suppose we have sampled from a normally distributed population with population variance unknown
- We can make use of the sample variance s^2: Estimate σ^2 with s^2
- Confidence intervals are now constructed as
  - $\overline{X} \pm z_{\alpha/2} \times s_{\overline{X}}$ when n is "large"
  - $\overline{X} \pm t_{\alpha/2,n-1} \times s_{\overline{X}}$ when n is "small"
- We need the t distribution because the sampling distribution of $\overline{X}$ is not quite normal.
- In the CIs above, $s_{\overline{X}} = \frac{s}{\sqrt{n}}$ and $t_{\alpha/2}$ has $n-1$ degrees of freedom.

The Student's t Distribution

- A random variable has a **Student's t distribution or t distribution** on ν degrees of freedom (or with parameter ν) if it can be expressed as the ratio of Z to $\sqrt{W/\nu}$ where $Z \sim N(0,1)$ and W (independent of Z) $\sim \chi^2_\nu$
- It can be shown that the construct $t = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ satisfies this condition.



The Student's t Distribution

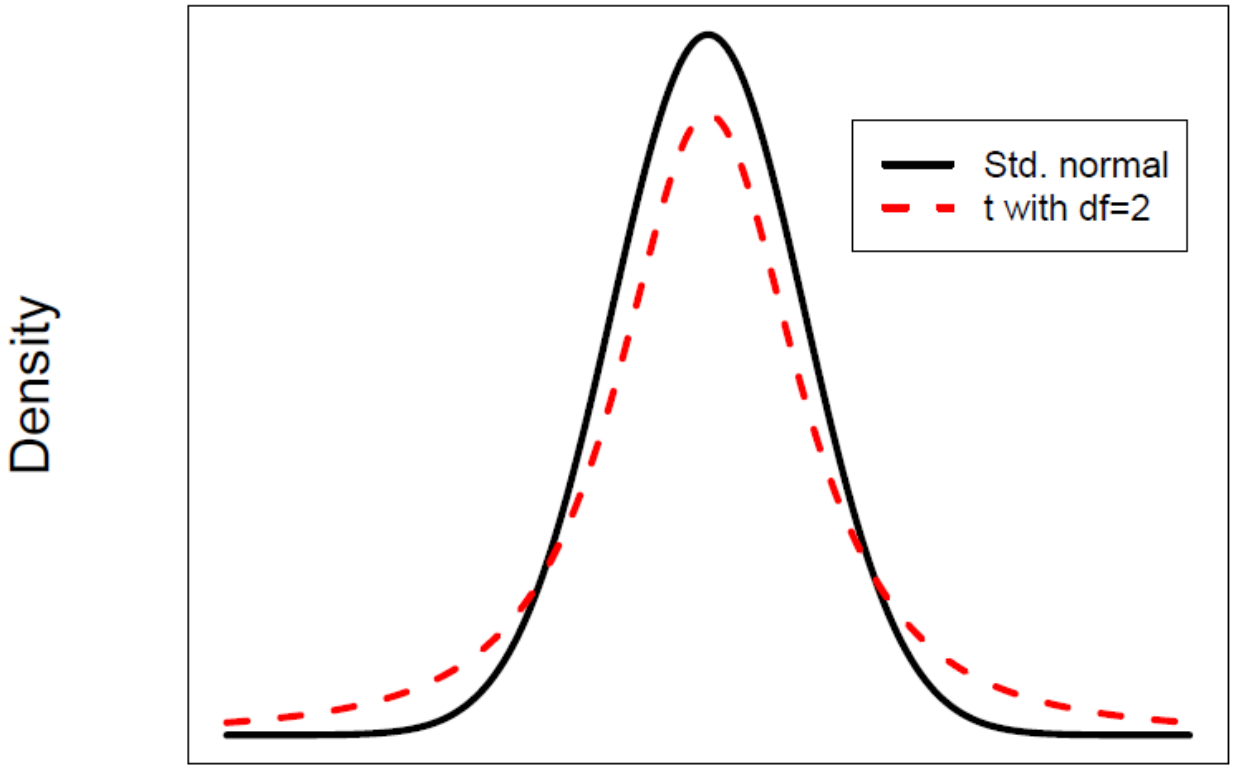
- If X is a random variable having density given by

$$f_X(x) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+x^2/k)^{(k+1)/2}},$$

  then X is defined to have a **Student's t distribution** or the density itself is called a *Student's t distribution with k degrees of freedom*

- Properties of $t = \frac{\overline{X}-\mu}{s/\sqrt{n}}$:
  - Symmetric about the mean, like the normal distribution
  - Mean = Median = Mode = 0
  - t ranges from $-\infty$ to $+\infty$
  - Encompasses a family of distributions determined by $\nu = n - 1$, the degrees of freedom
  - The t distribution approaches the standard normal distribution as $n - 1$ approaches ∞

Comparison of the Student's t with the Standard Normal



Like Z tables, there are T tables

Degrees of freedom, d	u								
	.75	.80	.85	.90	.95	.975	.99	.995	.9995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

*The uth percentile of a t distribution with d degrees of freedom.

Pivotal Quantity

- A **pivotal quantity or pivot** is generally defined as a function of observations and unobservable parameters whose probability distribution does not depend on unknown parameters
- Any probability statement of the form

$$P(a < H(X_1, X_2, \ldots, X_n; \theta) < b) = 1 - \alpha$$

  will give rise to a probability statement about θ
- Hence, pivots are crucial to construct confidence intervals for parameters of interest.
- Examples when sampling from a normal distribution:
  - $z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ (population variance known)
  - $t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$ (population variance unknown)

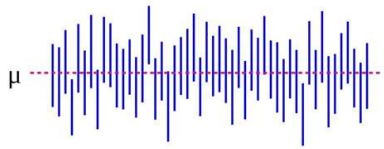
Confidence intervals for means

Summary table:

Population Distribution	Sample Size	Population Variance	95% Confidence Interval
Normal	Any	σ^2 known	$\bar{X} \pm 1.96\sigma/\sqrt{n}$
	Any	σ^2 unknown, use s^2	$\bar{X} \pm t_{0.025,n-1}s/\sqrt{n}$
Not Normal/ Unknown	Large	σ^2 known	$\bar{X} \pm 1.96\sigma/\sqrt{n}$
	Large	σ^2 unknown, use s^2	$\bar{X} \pm 1.96s/\sqrt{n}$
	Small	Any	Non-parametric methods
Binomial	Large	-	$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$
	Small	-	Exact methods

Interpretation of Confidence Interval (CI)

- *Before* the data are observed, the probability is at least $(1-\alpha)$ that $[L, U]$ will contain the population parameter
- In *repeated sampling* from the relevant distribution, $100(1-\alpha)\%$ of all intervals of the form $[L, U]$ will include the true population parameter



- *After* the data are observed, the constructed interval $[L, U]$ either contains the true parameter value or it does not (there is no longer a probability involved here!)

A statement such as $P(3.5 < \mu < 4.9) = 0.95$ is **incorrect** and should be replaced by **A 95% confidence interval for μ is (3.5,4.9)**