STAT 346/446 Lecture 12

Loss-function optimality – Decision Theory

CB Sections 7.3.4 and 8.3.5, DS Sections 7.4 and 9.8

Decision Theory (loss-function optimality)

- "Best" point estimators
- "Best" hypothesis testing procedure
- MSE is a special case of a *loss function*
- After we observe data $\mathbf{X} = \mathbf{x}$ a decision has to be made about θ
- Decision theory: Take into account the losses that can occur when making a decision about θ

Decision Theory

- **Sample space** S: Set of all possible samples (observations).
  - We observe $\mathbf{X} = \mathbf{x}$ where $\mathbf{x} \in S$
- **Parameter space** Θ: Set of possible values for the unknown parameters θ.
  - The unknown *true* value of the parameter is in Θ
  - Sometimes called *states of nature*
- **Action space** $\mathcal{A}$: The set of all possible actions the statistician can take.
  - Also called *decision space* and *decisions*
  - Sometimes $\mathcal{A} = \Theta$
- **Action rule** is a function δ from S into $\mathcal{A}$.
  - The action we take depends on the action rule: $a = \delta(\mathbf{x})$
  - The action rule is usually some statistical procedure and δ can be a statistic

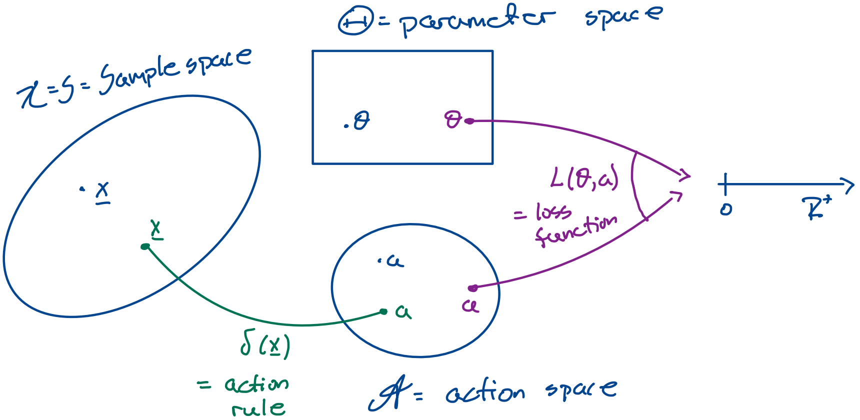
Point estimation as a decision problem

- Sample space S: Set of all possible samples
  - We observe $\mathbf{X} = \mathbf{x}$ where $\mathbf{x} \in S$
- States of nature Θ: Parameter space
  - Set of all possible values of the true value of the parameter θ
- Action space: Here we have $\mathcal{A} = \Theta$
  - Action: the estimate we come up with for θ
- Action rule: Point estimator, e.g. $\delta(\mathbf{x}) = \overline{x}$

Hypothesis testing as a decision problem

- Sample space S: Set of all possible samples
  - We observe $\mathbf{X} = \mathbf{x}$ where $\mathbf{x} \in S$
- States of nature Θ: Parameter space
  - Set of all possible values of the true value of the parameter θ
- Action space: $\{a_0, a_1\} = \{\text{don't reject } H_0, \text{reject } H_0\}$
  - Action: we choose between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$
- Action rule: Decision rule
  - e.g. choose H_1 if $\mathbf{x} \in R$ and choose H_0 if $\mathbf{x} \in R^c$

$$\delta(\mathbf{x}) = \begin{cases} \text{reject } H_0 & \text{if } \mathbf{x} \in R \\ \text{don't reject } H_0 & \text{if } \mathbf{x} \notin R \end{cases}$$



Risk function: $R_\delta(\theta) = E\left(L(\theta, \delta(\underline{X}))\right) = \int_{\mathcal{X}} L(\theta, \delta(\underline{x}))\, f(\underline{x})\, d\underline{x}$

||

Expected loss   book: $R(\theta, \delta(\underline{x}))$

Loss functions for point estimation

S: Sample space Θ: Parameter space $\mathcal{A}$: Action space

- **Loss function** $L : \Theta \times \mathcal{A} \to \mathbb{R}^+$
  - $L(\theta, a) =$ The loss when the true value of the parameter is θ and action a is taken

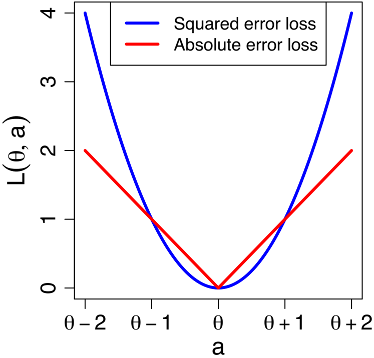
- Examples of loss functions in point estimation problems:

- **Squared error loss**:

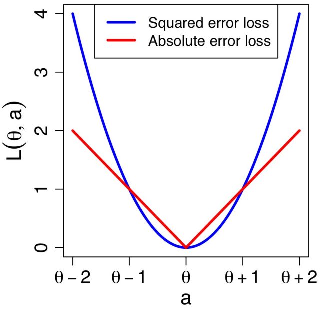
$$L(\theta, a) = (\theta - a)^2$$

- **Absolute error loss**:

$$L(\theta, a) = |\theta - a|$$



θ fixed



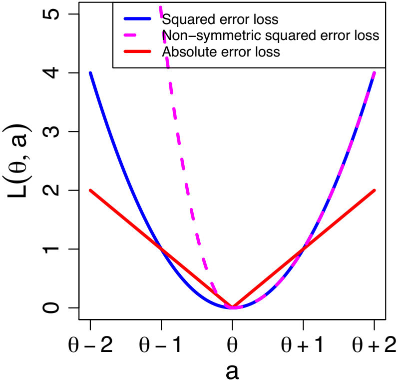
Notes

* If $a=\theta$ (correct action) $L(\theta,a)=0$ for both loss functions
  → No loss for correct action
* Both are symmetric, so over and under estimation are penalized the same way
* Both functions have losses that increase when the action is further away from the true value of θ
  - absolute loss: linear increase with distance
  - sq. error loss: small loss for small differences but very high loss for large differences

Non-symmetric squared error loss function

- Both absolute and squared error losses penalize over and under estimation equally
- Example of a loss function that penalizes under estimation more than over estimation:

$$L(\theta, a) = \begin{cases} 5(\theta - a)^2 & \text{if } a < \theta \\ (\theta - a)^2 & \text{if } a \geq \theta \end{cases}$$



Risk Function

- The loss function depends on the sample through the action rule:

$$L(\theta, a) = L(\theta, \delta(\mathbf{x}))$$

$\leftarrow$ function of $\underline{x}$

$L(\theta, \delta(\underline{X}))$ is a rand variable

- **Risk function** of a statistical procedure $\delta(\mathbf{X})$ is the expected loss:

$$R_\delta(\theta) = R(\theta, \delta) = \mathrm{E}_\theta\left(L(\theta, \delta(\mathbf{X}))\right) = \int_{\mathcal{X}} L(\theta, \delta(\underline{x}))\, f(\underline{x} | \theta)\, d\underline{x}$$

$\hookrightarrow$ function of θ. Get different risk functions for different δ (action rule)

E_θ is the expectation with respect to $\mathbf{X}$ for fixed θ.

- Goal: To find a statistical procedure with minimum risk
    - Complicated by the fact that $R(\theta, \delta)$ usually depends on θ (the actual state of nature)
    - Ideally our estimator has smallest risk for all θ

Risk functions for point estimation

- For squared error loss the risk is the mean squared error:

$$R(\theta, \delta) = \mathrm{E}_\theta\left((\delta(\mathbf{X}) - \theta)^2\right) = \mathrm{Var}(\delta(\mathbf{X})) + \mathrm{bias}(\delta(\mathbf{X}))^2$$

- For absolute error loss the risk is

$$R(\theta, \delta) = \mathrm{E}_\theta\left(|\delta(\mathbf{X}) - \theta|\right)$$

So, minimizing MSE for a point estimator is the same as minimizing risk under squared error loss.

Example

- $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$
- Consider squared error loss and three estimators for σ^2

$$S^2, \qquad \hat{\sigma}^2 = \frac{n-1}{n} S^2 \qquad \text{and} \qquad \tilde{S}^2 = \frac{n-1}{n+1} S^2$$

Found last time:

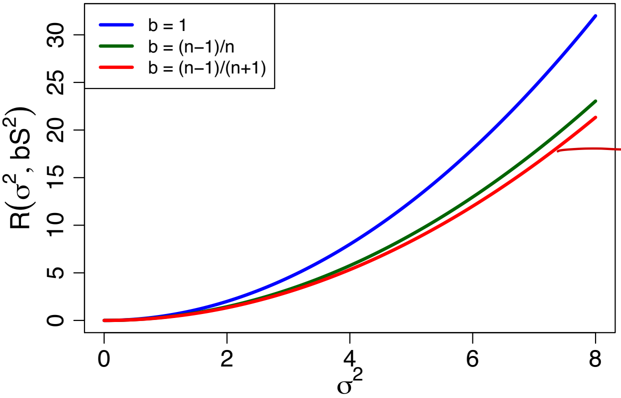
$$R(\sigma^2, bS^2) = \left(\frac{2b^2}{n-1} - (b-1)^2 \right) \sigma^4$$

$\to$ minimized at $b = \frac{n-1}{n+1}$

Example

Same as Figure 7.3.2 in the book

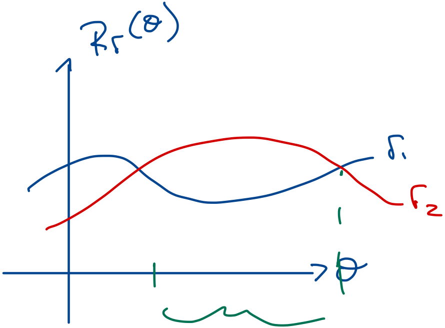
Considering only estimators of the form bS^2



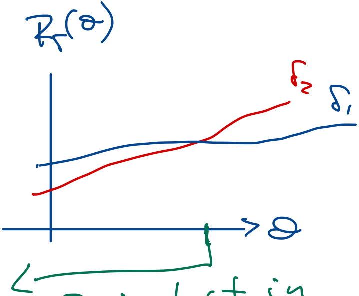
i.e. over all σ^2

uniformly smallest risk

Often there is is not a uniformly lowest risk action rule



δ_1 is best for θ in this range



δ_2 is best in this range

no overall winner

Bayes risk

- Risk is usually a function of θ
- Perhaps we care more about some values of θ than others
  - Expressed via the prior distribution on θ: $\pi(\theta)$
- **Bayes risk** is defined as

$$E(R_\delta(\theta)) = \int_\Theta R(\theta,\delta)\pi(\theta)d\theta$$

Gives *one number* for each action rule, so easier to compare.

Bayes risk and Posterior expected loss

- We can (for most distributions) switch the order of integration:

posterior

$$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \underbrace{\left(\int_\mathcal{X} L(\theta,\delta(\mathbf{x}))f(\mathbf{x}\mid\theta)d\mathbf{x}\right)}_{R_\delta(\theta)}\pi(\theta)d\theta$$

$$= \int_\mathcal{X}\left(\int_\Theta L(\theta,\delta(\mathbf{x}))\underbrace{f(\mathbf{x}\mid\theta)\pi(\theta)}_{*}d\theta\right)d\mathbf{x}$$

$$= \int_\mathcal{X}\left(\int_\Theta L(\theta,\delta(\mathbf{x}))\overbrace{\pi(\theta\mid\mathbf{x})m(\mathbf{x})}d\theta\right)d\mathbf{x}$$

$$= \int_\mathcal{X}\underbrace{\left(\int_\Theta L(\theta,\delta(\mathbf{x}))\pi(\theta\mid\mathbf{x})d\theta\right)}_{\text{posterior expected loss}} m(\mathbf{x})d\mathbf{x}$$

$$* \quad \pi(\theta \mid \underline{x}) = \frac{f(\underline{x}\mid\theta)\pi(\theta)}{m(\underline{x})}$$

$$\Rightarrow f(\underline{x}\mid\theta)\pi(\theta) = \pi(\theta\mid\underline{x})\, m(\underline{x})$$

Posterior expected loss

- **Posterior expected loss** is defined as

$$\int_\Theta L(\theta, \delta(\mathbf{x}))\pi(\theta \mid \mathbf{x})d\theta$$

← a function of $\underline{x}$

- **Bayes rule** = the action rule that minimizes the posterior expected loss for any observed sample **x**.

- For squared error loss we get

post. exp. loss:

posterior

min for $E(\theta / X=x)$

$$\int_\Theta (\theta - \delta(\mathbf{x}))^2\pi(\theta \mid \mathbf{x})d\theta = E\left((\theta - \delta(\mathbf{x}))^2 \mid \mathbf{X} = \mathbf{x}\right)$$

Bayes rule: $E(\theta \mid \mathbf{x})$ = posterior mean ← Reason we call the post. mean the Bayes estimator

- Bayes rule for absolute error loss is the posterior median

- **Bayes rule** = the action rule that minimizes the posterior expected loss for any observed sample **x**.

The point is: If we can find the δ with the smallest posterior expected loss for all $\underline{x} \in \mathcal{X}$, then we have found the δ with smallest Bayes risk.

(i.e. best action rule w.r.t the chosen loss and prior)

In general: Let Y be a random variable

The constant c that minimizes

$$E((Y-c)^2) \quad \text{is} \quad c = E(Y)$$

The constant c that minimizes

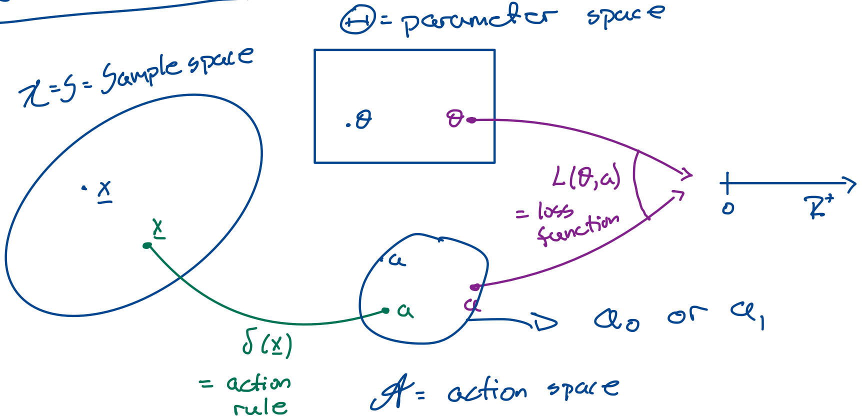
$$E(|Y-c|) \quad \text{is} \quad c = \text{median}(Y)$$

Note: We get a much more general result for point estimation in Bayesian inference than frequentist.

* Under squared error loss, posterior mean is the best estimator

* Under absolute error loss, posterior median is the best estimator

Hypothesis testing



Risk function: $R_\delta(\theta) = E\big(L(\theta, \delta(\underline{X}))\big) = \int_{\mathcal{X}} L(\theta, \delta(\underline{x})) f(\underline{x})\, d\underline{x}$

$\shortparallel$

Expected loss

book: $R(\theta, \delta(\underline{x}))$

Loss functions for hypothesis testing

S: Sample space Θ: Parameter space $\mathcal{A}$: Action space

$= \{a_0, a_1\}$

- **Loss function** $L : \Theta \times \mathcal{A} \to \mathbb{R}^+$
  - $L(\theta, a) =$ The loss when the state of nature is θ and action a is taken
- In hypothesis testing there are only two possible actions and two relevant states of nature

	$\theta \in \Theta_0$	$\theta \in \Theta_0^c$
a_0: choose H_0		Type II error
a_1: choose H_1	Type I error	

Loss functions for hypothesis testing

loss for type I and II errors the same.

- **0-1 loss**:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ 1 & \text{if } \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \\ 0 & \text{if } \theta \in \Theta_0^c \end{cases}$$

$L(\theta, a_i)$	$\theta \in \Theta_0$	$\theta \in \Theta_0^c$
a_0	0	1 *
a_1	1 *	0

* could have any number here, only matters if equal or not

function of $\underline{X}$

$$L(\theta, a) = \begin{cases} 0 & \text{if } (a = a_0 \text{ and } \theta \in \Theta_0) \text{ or } (a = a_1 \text{ and } \theta \in \Theta_0^c) \\ 1 & \text{if } (a = a_0 \text{ and } \theta \in \Theta_0^c) \text{ or } (a = a_1 \text{ and } \theta \in \Theta_0) \end{cases}$$

random variable

Loss functions for hypothesis testing

- **Generalized 0-1 loss**:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ c_2 & \text{if } \theta \in \Theta_0^c \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} c_1 & \text{if } \theta \in \Theta_0 \\ 0 & \text{if } \theta \in \Theta_0^c \end{cases}$$

$L(\theta, a_i)$	H_0 is true $\theta \in \Theta_0$	H_0 is false $\theta \in \Theta_0^c$
a_0	0	c_1
a_1	c_2	0

$\frac{c_1}{c_2}$ is the important thing, not the values of c_1 and c_2

$$L(\theta, a) = \begin{cases} 0 & \text{if } (a = a_0 \text{ and } \theta \in \Theta_0) \text{ and } (a = a_1 \text{ and } \theta \in \Theta_0^c) \\ c_1 & \text{if } a = a_0 \text{ and } \theta \in \Theta_0^c \\ c_2 & \text{if } a = a_1 \text{ and } \theta \in \Theta_0 \end{cases}$$

Risk function for hypothesis tests

- Under the generalized 0-1 loss, the risk function is closely related to the power function
- Action rule:

rejection region

$$\delta(\mathbf{x}) = \begin{cases} a_1 & \text{if } \mathbf{x} \in R \\ a_0 & \text{if } \mathbf{x} \in R^c \end{cases}$$

- Risk function:

discr. random var. that can take values 0, c_1, or c_2

$$R_\delta(\theta) = E\Big(\underbrace{L(\theta, \delta(\underline{X}))}\Big)$$

$$= 0\, P(L(\theta,\delta(\underline{X})) = 0) + c_1 P_\theta(L(\theta,\delta(\underline{X})) = c_1) + c_2 P_\theta(L(\theta,\delta(\underline{X})) = c_2)$$

note: $P(L(\theta,\delta(\underline{x})) = c_1) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ P(a_0) & \text{o.w.} \end{cases}$

$$R_\delta(\theta) = c_1 P(L(\theta,\delta(\underline{X})) = c_1) + c_2 \underbrace{P(L(\theta,\delta(\underline{X})) = c_2)}_{=0 \text{ if } \theta \in \Theta_0^c}$$

For $\theta \in \Theta_0^c$:

$$R_\delta(\theta) = c_1 P(\delta(\underline{X}) = a_0) + c_2 \cdot 0$$

$$= c_1 P(\text{don't reject}) = c_1 (1 - P(\text{reject}))$$

$$= c_1 (1 - \beta(\theta))$$

$\beta(\theta)$ = power function

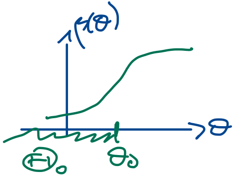
For $\theta \in \Theta_0$:

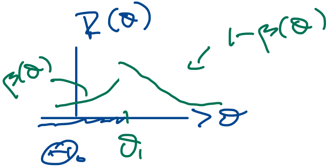
$$R_\delta(\theta) = c_1 \cdot 0 + c_2 P(\delta(\underline{X}) = a_1)$$

$$= c_2 P(\text{reject}) = c_2 \beta(\theta)$$

Risk function can weigh loss with the power function

$c_1 = c_2 = 1$





want $R(\theta)$ small everywhere

Minimizing risk w.r.t. 0-1 loss
is equivalent to minimizing probabilities
of type I and type II errors

i.e. setting sign. level low and
finding the most powerful test.

Generalized 0-1 loss: Can weigh
the importance of type I and
type II errors differently.

Example

Exercise 8.56 from textbook

$\Theta_0 = (0, 1/3]$ $\Theta_0^c = (1/3, 1]$

- $X \sim \text{Binomial}(5, p)$. Want to test $H_0 : p \leq 1/3$ versus $H_1 : p > 1/3$, using 0-1 loss $(c_1 = c_2 = 1)$
- Compare Risk function for two test procedures:
  - δ_1 rejects if $X = 0$ or $X = 1$ ← a rather silly test.
  - δ_2 rejects if $X = 4$ or $X = 5$

Risk function for δ_1 : $= \beta(p)$

for $p \leq \frac{1}{3}$ (Θ_0). $R_{\delta_1}(p) = P(\text{reject}) = P(X=0 \text{ or } X=1)$

$= \binom{5}{0} p^0 (1-p)^5 + \binom{5}{1} p (1-p)^4$

for $p > \frac{1}{3}$ (Θ_0^c): $R_{\delta_1}(p) = 1 - P(\text{reject}) = 1 - (1-p)^5 + 5p(1-p)^4$

$\quad = 1 - \beta(p)$

Risk function for test δ_2:

For $p \le 1/3$: $R_{\delta_2}(\theta) = \beta(p) = P(\text{reject } H_0)$

$R(\theta, \delta_2)$

$$= P(X=4 \text{ or } X=5)$$

$$= \binom{5}{4} p^4 (1-p) + \binom{5}{5} p^5 (1-p)^0$$

$$= 5p^4(1-p) + p^5$$

For $p > \frac{1}{3}$: $R_{\delta_2}(\theta) = 1 - \beta(p) = 1 - P(\text{reject})$

$$= 1 - 5p^4(1-p) + p^5$$

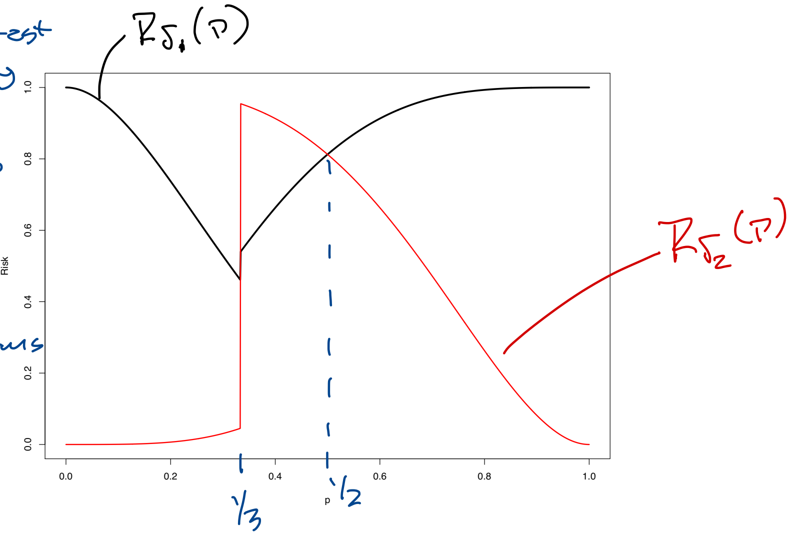
Example - Risk functions

want minimum risk

Neither test is uniformly better.

δ_1 performs better for $p \in (\frac{1}{3}, \frac{1}{2})$

Otherwise δ_2 performs better.



Example - R code for Risk functions

```r
R1 <- function(p,p0){
  tmp <- (1-p)^5 + 5*p*(1-p)^4
  tmp[p>p0] <- 1 - tmp[p>p0]
  return(tmp)
}

R2 <- function(p,p0){
  tmp <- p^5 + 5*p^4*(1-p)
  tmp[p>p0] <- 1 - tmp[p>p0]
  return(tmp)
}

curve(R1(x, p0=1/3), from = 0, to = 1, ylim=c(0,1), n=1001, lwd=3,
      ylab="Risk", xlab="p")
curve(R2(x, p0=1/3), from = 0, to = 1, col='red', add=T, n=1001, lwd=2)
```

Example

Exercise 8.55 from textbook

- $X \sim N(\theta, 1)$, want to test $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, using the following loss function:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \geq \theta_0 \\ b(\theta_0 - \theta) & \text{if } \theta < \theta_0 \end{cases} \quad \text{and}$$

c_1

$$L(\theta, a_1) = \begin{cases} c(\theta_0 - \theta)^2 & \text{if } \theta \geq \theta_0 \\ 0 & \text{if } \theta < \theta_0 \end{cases}$$

c_2

- Test procedures: Reject if $X < -z_\alpha + \theta_0$ for $\alpha = 0.1, 0.3, 0.5$
- Find the risk function and compare

3 test procedures.

Risk function for test δ_α:

For $\theta \geq \theta_0$: $R_{\delta_\alpha}(\theta) = c(\theta_0 - \theta)^2 P(\text{reject})$

Θ_0

$$= c(\theta_0 - \theta)^2 P(X < -z_\alpha + \theta_0)$$

$X \sim N(\theta, 1)$

$$= \underbrace{c(\theta_0 - \theta)^2}\ \underbrace{\Phi(-z_\alpha + \theta_0 - \theta)}$$

functions of θ

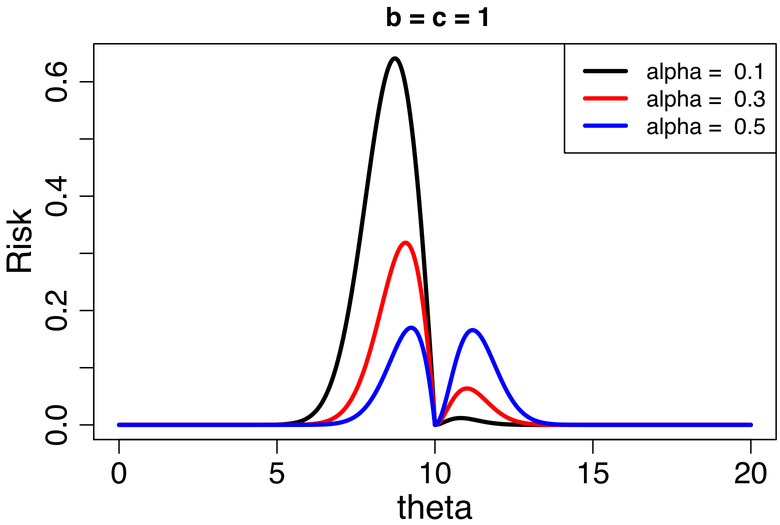
For $\theta < \theta_0$: $R_{\delta_\alpha}(\theta) = b(\theta_0 - \theta)(1 - P(\text{reject}))$

$$= b(\theta_0 - \theta)(1 - P(X < -z_\alpha + \theta_0))$$

$$= b(\theta_0 - \theta)(1 - \Phi(-z_\alpha + \theta_0 - \theta))$$

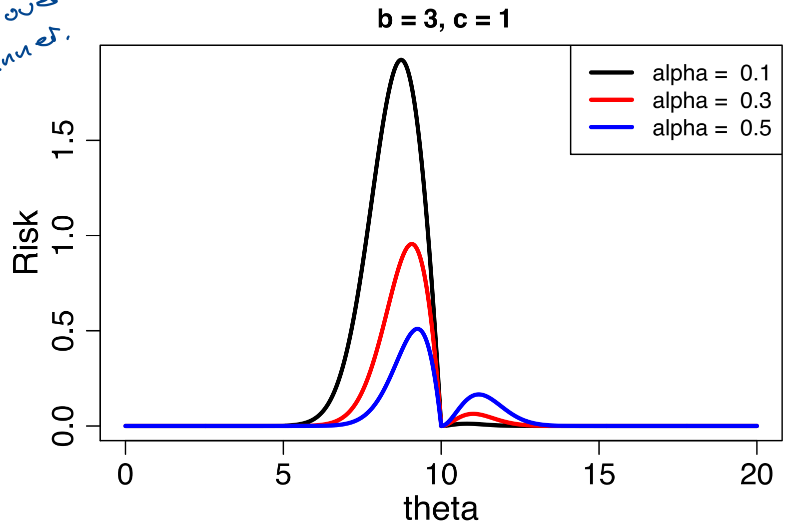
Example - Risk functions

$\theta_0 = 10$

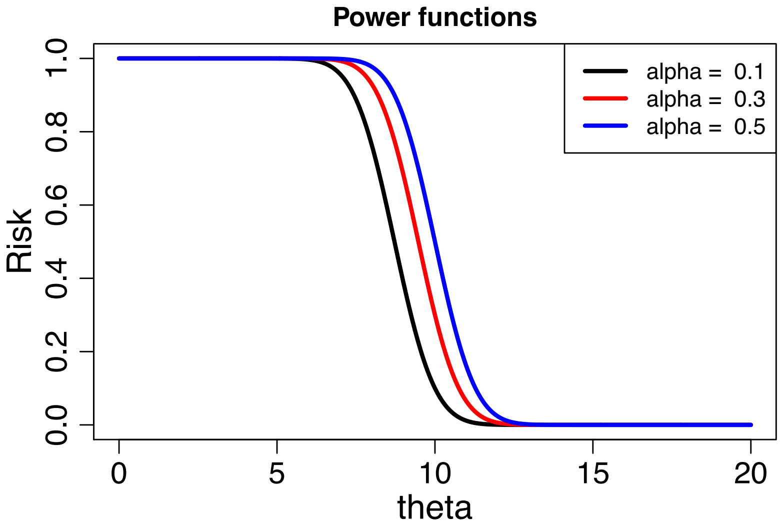


Example - Risk functions

no overall winner.



Example - Risk functions



Point: When comparing tests we rely on the power function

Risk function is a combination of the power function and loss function.