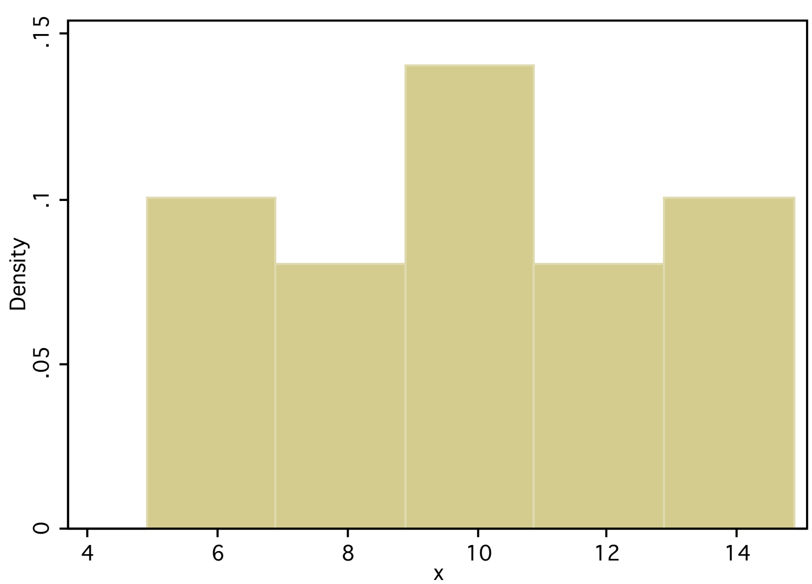
Lecture #2: Efficiency and Robustness

September 29, 2009
Biostatistics 210

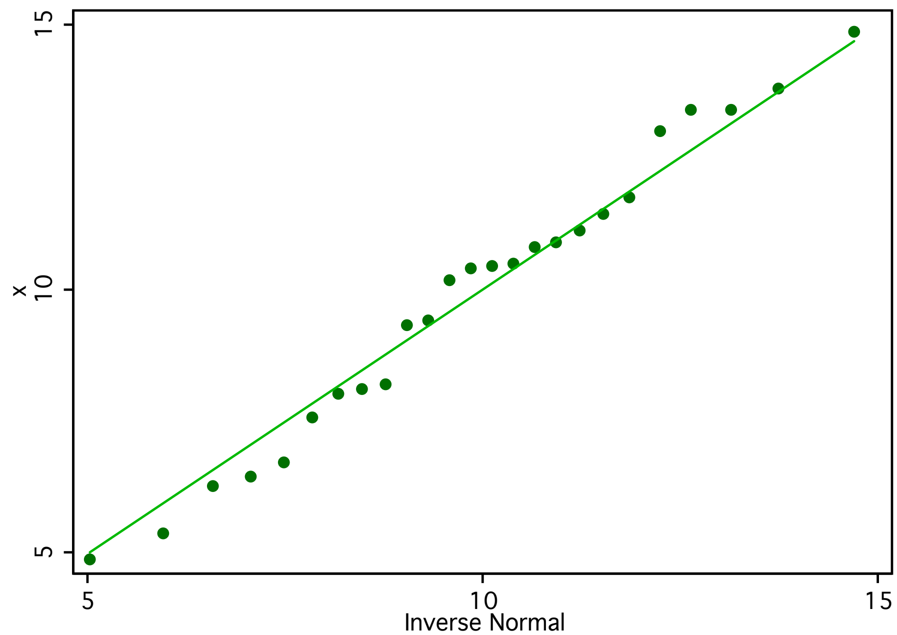
Small Dataset

n=25



Should I use the mean or median?

Normal Quantile Plot



So Mean or Median

- How to decide which is better
- Data appears Normal -- should estimate the same thing
- Which gives a "better" estimate
- How can we define "better"

Summary

- Statistical estimation is like target practice
- Want to get close to target
    - by choosing right technique
- Two kinds of errors: systematic (bias) and random (variability)

Unbiased Estimator

My shooting was correct "on average"

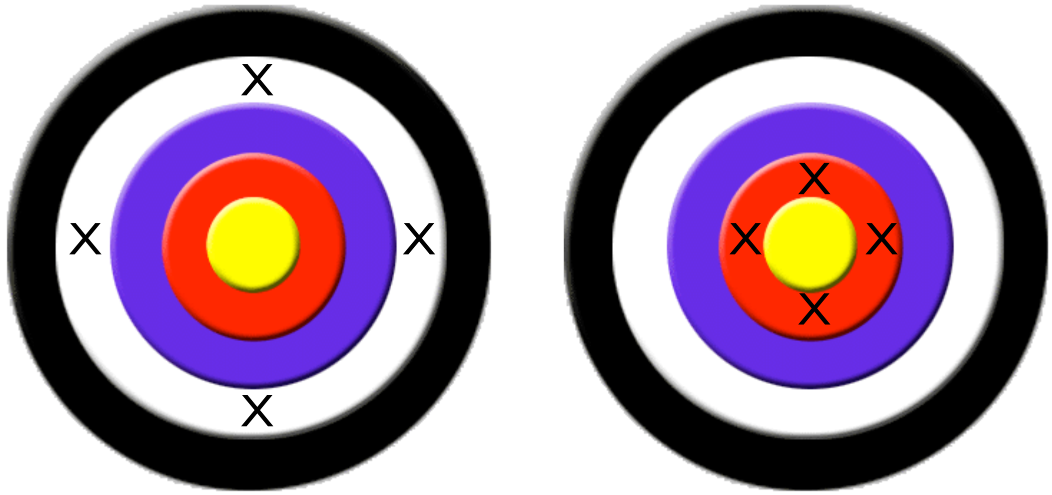
Biased Estimator

Tendency to shoot above and to right of target

Target is good analogy

- Try to get close to truth:
  minimize avg distance betw' estimator & truth
- Distance = $(\text{bias})^2$ + variance of estimator
- var(estimator) = $(\text{standard error squared})^2$
  minimize avg distance betw' estimator & truth

Two Unbiased Estimators



Two Estimators

- Both unbiased: no systematic errors
- Only random errors
- Errors consistently smaller on right
- That is *efficiency*

Efficiency

When comparing estimators, an estimator is more efficient if it has a variance that is smaller compared to others. In some cases, there is an estimator with the lowest variance. Such an estimator would be said to be the most efficient.

We only care about efficiency of estimators that are unbiased or are consistent

Here's why

Small variance, bad estimator

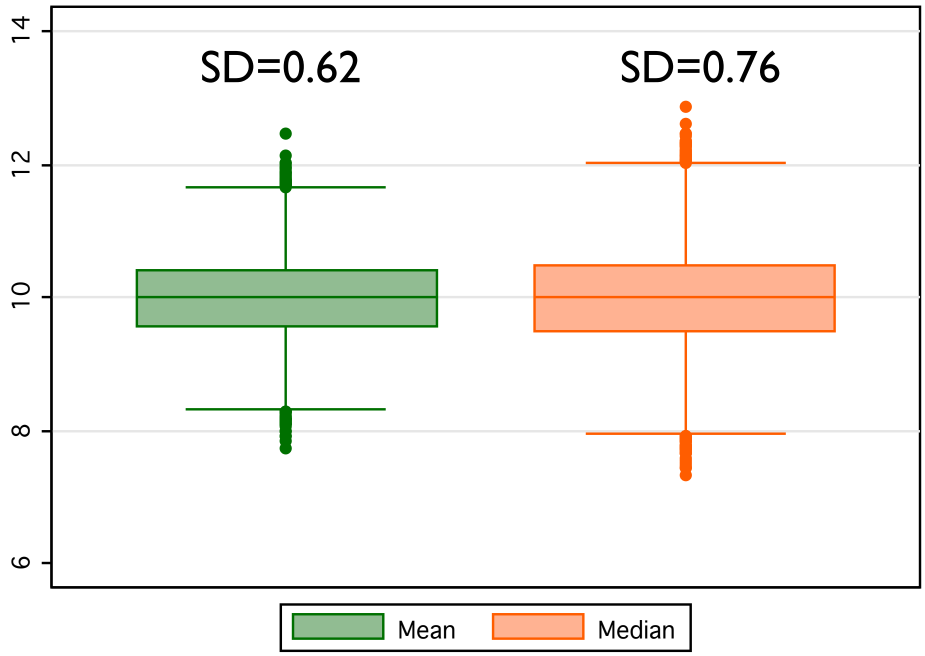
Application to Data

- Specific Question: Which is more efficient: the mean or the median?
- Answer may depend on data features
- Our data: n=25, approximately Normal
- Can figure out mathematically
- Or use simulation

Simulation

- N=25
- Y ~ Normal (mu=10, sigma=3)
- Mean(Y) = Median(Y) = 10
- Is mean or median closer to mu on average?

10,000 Mean and Medians



Mean gets closer

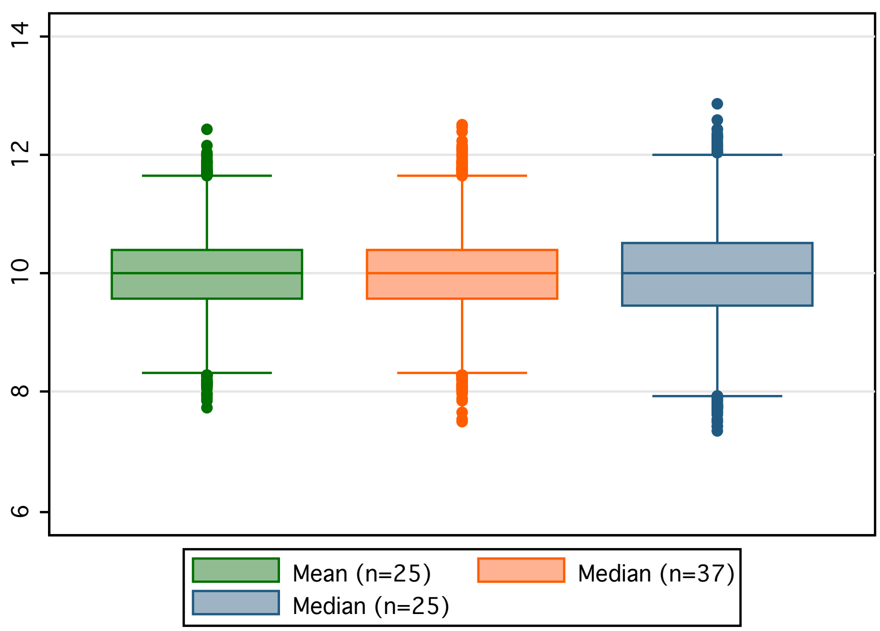
- More efficient
- Why do you think that is?

Relative Efficiency

$$\frac{SE(median)^2}{SE(mean)^2} = \frac{(0.76)^2}{(0.62)^2} = 1.50$$

mean based on 25 is as accurate as median calculated on RE x 25 people = 37 people

Graphically



Consider

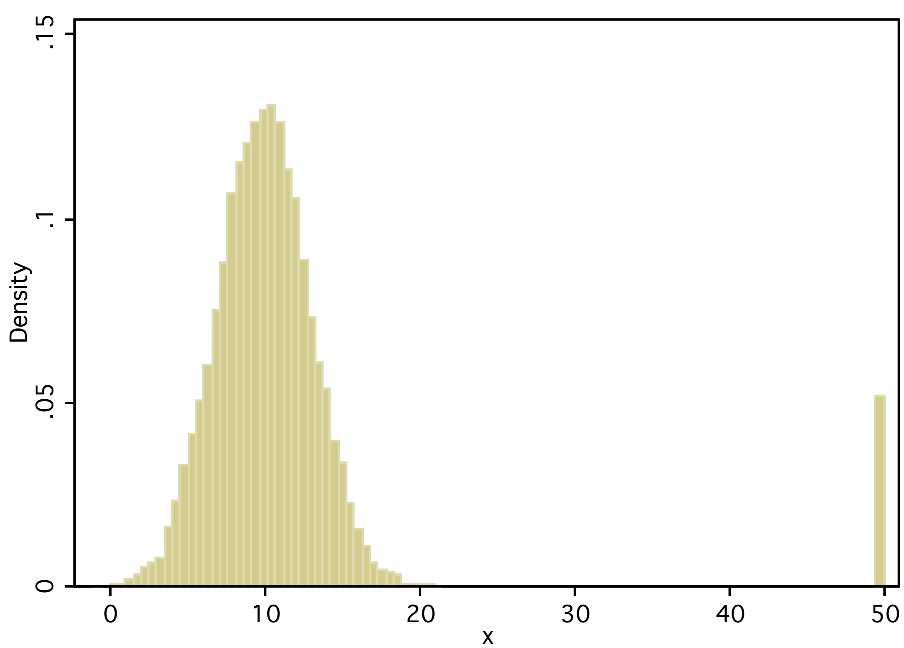
- N=500
- Y ~ Normal(mu=10, sigma =3)
- calculate mean and median
- 10,000 simulations
- Rel efficiency = 1.58
  doesn't really depend on n

Should we always use the mean over the median?

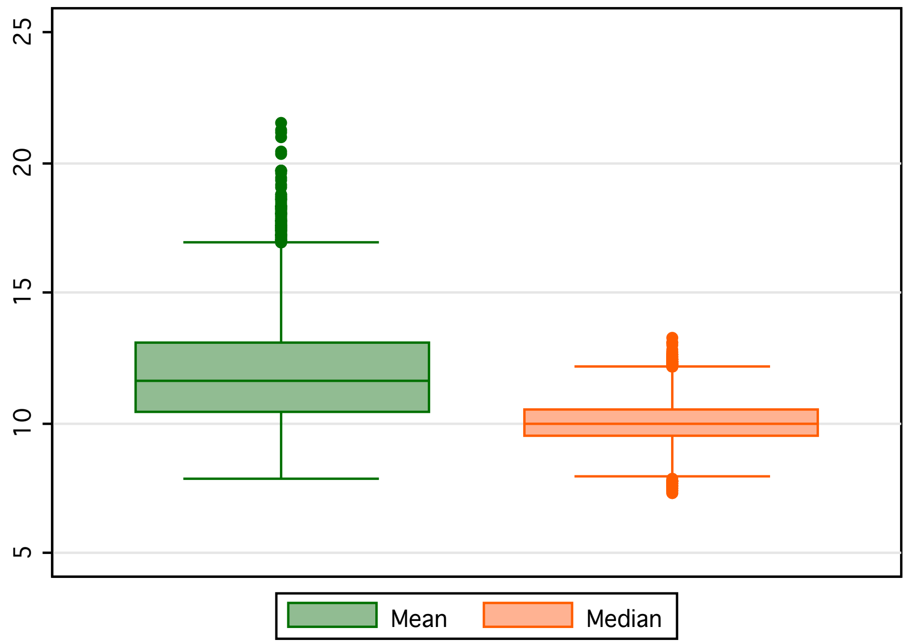
What if

- Data is so close to a Normal distribution with mu=10, sigma=3
- But there is a small (5%) chance of outliers
- A HUGE one -- 50
- How would this change our preference?

Data Density



Mean v. Median



Why is that?
So What?

Why?

- The mean is not robust
- Uses all the data:
  doesn't downweight obvious outliers
- "It's not a bug, it's a feature"
- mean optimal for Normal data
  can't tolerate outliers well

So What?

- Efficient estimators are derived under assumptions
- Can underperform when assumption violated
- Robust estimators are more resilient
- More robust => less efficient
- More efficient => less robust

Trimmed Mean

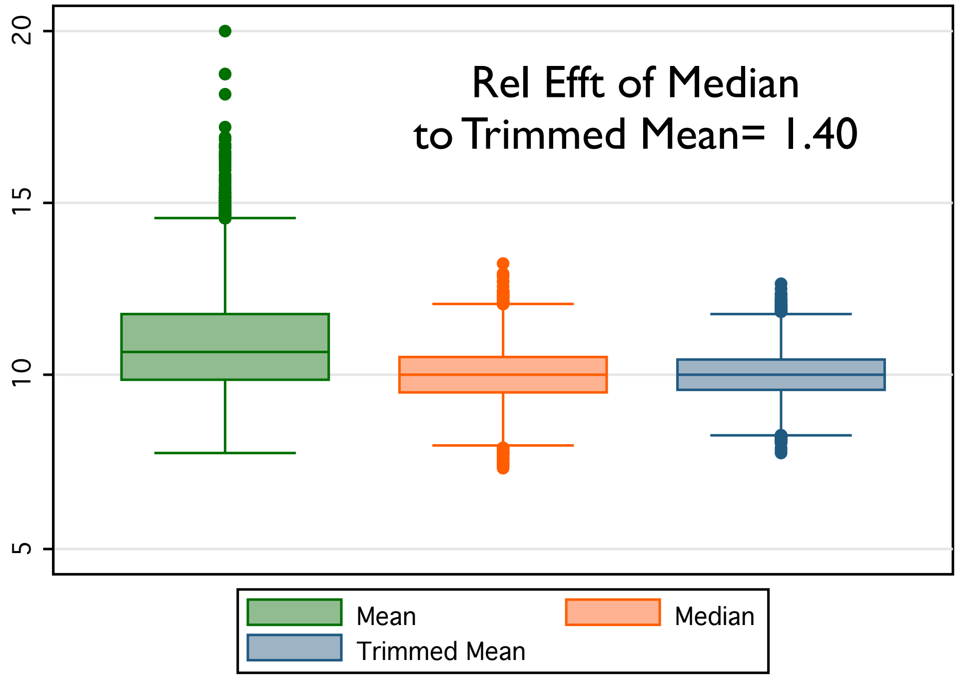
robust version of the mean

- 10% Trimmed Mean:
  1. drop lower and upper 10% of data
  2. calculate mean
- Discards outliers
- More cautious than mean
- Uses more data than median

How does it perform

- Consider the n=25
- Normal data with a slight contamination with extreme outliers
- Simulate 10,000 datasets
- Compare: mean, median and trimmed mean

Results

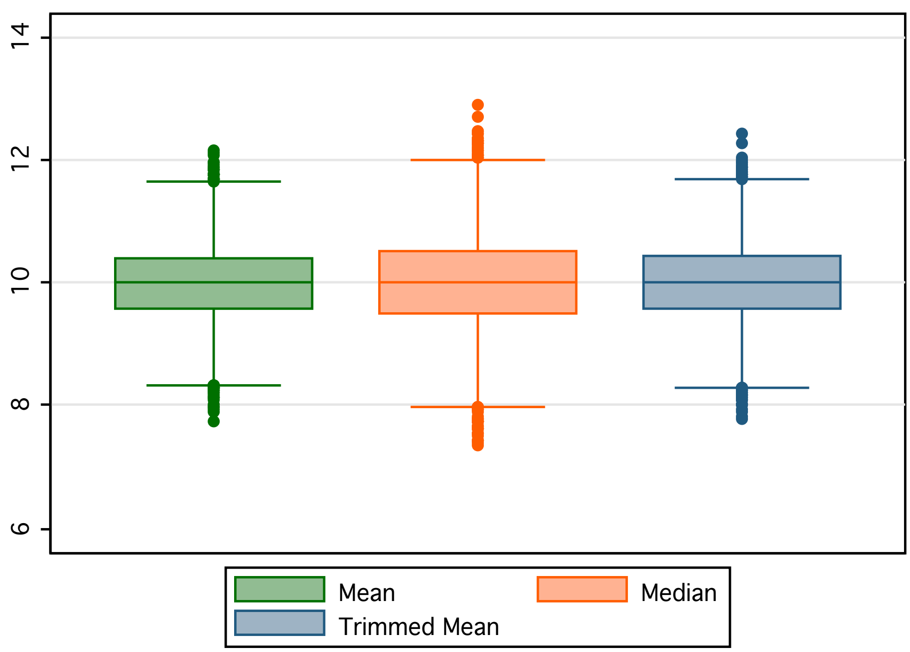


Relative MSE

Statistic	Mean	Mean Squared Error	Relative MSE
Mean	10.9	2.80	6.41
Median	10.0	0.61	1.40
Trimmed Mean	10.1	0.44	---

What if data is Normal?

Boxplots of Statistics



Relative Efficiency

Statistic	Mean	Standard Error	Relative Efficiency
Mean	10.0	0.61	-----
Median	10.0	0.76	1.55
Trimmed Mean	10.0	0.63	1.07

Trimmed Mean

- Less robust but more efficient than the median
- More robust but less efficient than the mean
- Mean: 0% trimmed mean
- Median 49.9% trimmed mean

Bottom Line

- Judge point estimates by
  - validity (unbiasedness/consistency)
  - efficiency
  - robustness
- Tension between efficiency and robustness

Hypothesis Testing

- Two groups of subjects
- n1=12, n2=38
- Continuous data
- H_0: $\mu_0 = \mu_1$
- Should I use a equal or unequal variance test?
- How can we discuss this choice

In terms of size and power

Hypothesis Testing

- α: type 1 error
  almost always set to be 0.05 or 0.025
- β: type 2 error
- α: probability test is significant under H_0
  e.g., probability t-test sig if means equal
- $1-\beta$: probability test significant under H_1
  the power

Comparing Tests

- Validity: pr test rejects = 0.05 under H_0
  probability $p < 0.05$ is 0.05
- Suppose I arbitrarily reject with prob 0.05
  ignoring the data
- Would that give a valid test?
- But....also want power

Data Simulation

unequal variance, H_0 true

- n_0=20, n_1=50
- group 0: Y ~ Normal(μ_0 =10, sigma=3)
- group 1: Y ~ Normal(μ_1=10, sigma=5)
- H_0: $\mu_0 = \mu_1$ (but variances unequal)
- H_0 is true for this data
- How often do two t-tests reject

Test "Size"

true variance unequal

Test	# Rejecting	Proportion Rejected
t-test (eq)	189	0.02
t-test (uneq)	532	0.05

Note, type I errors less than 0.05 or eq var test

Data Simulation

equal variance

- n_0=20, n_1=50
- group 0: Y ~ Normal(μ_0 =10, sigma=3)
- group 1: Y ~ Normal(μ_1 =12, sigma=3)
- equal variances
- $\delta = \mu_0 - \mu_1$
  measure of how false H_0 is
- Power: equal and unequal variances

Power Simulations

equal variances

Test	# Rejecting	Proportion Rejected
t-test (eq)	7016	0.70
t-test (uneq)	6887	0.68

Both tests have well-controlled type 1 error

Data Simulation

unequal variance, H_0 false

- n_0=20, n_1=50
- group 0: Y ~ Normal(μ_0 =10, sigma=3)
- group 1: Y ~ Normal(μ_1 =12.7, sigma=5)
- $\delta = \mu_0 - \mu_1$
  measure of how false H_0 is
- Power: equal and unequal variance test

Test Power

equal variances, H_0 false

Test	# Rejecting	Proportion Rejected
t-test (eq)	6275	0.63
t-test (uneq)	7774	0.78

Both tests have well-controlled type 1 error

Summary

test	Are the true variances?			
	Equal		Unequal	
	Size	Power	Size	Power
Equal var	0.05	0.70	0.02	0.63
Unequal	0.05	0.68	0.05	0.78

Summary

test	Are the true variances?			
	Equal		Unequal	
	Valid	Efft	Valid	Efft
Equal var	Y	Y	Y*	Y
Unequal	Y	Y	Y	N

**test size < 0.05 is technically "valid"*

Choosing t-test

- Both are valid (0.05 level test) under their assumption
- If variance unequal, = variance test has < 0.05 size and quite a bit less power
- If variance equal, ≠variance has 0.05 size and only slightly power
- Choice between them is clear

Caveats

- Assumes data is Normal
  with outliers, Wilcoxon is better
- Has unequal n_0 and n_1
  if equal, differences less dramatic
- But, illustrates how we can discuss choice

Major Idea

- We need to choose between methods
- Helpful to consider statistical properties
  validity, efficiency, robustness
- Practical properties: *simplicity, interpretation*
- Leads to a more productive discussion