Help yourself to a cupcake!

1/3 thanks to everyone who bought a textbook on amazon

Stat310

Fin

Hadley Wickham

1. Why statistics?
2. Stats @ Rice
3. Review
4. Feedback

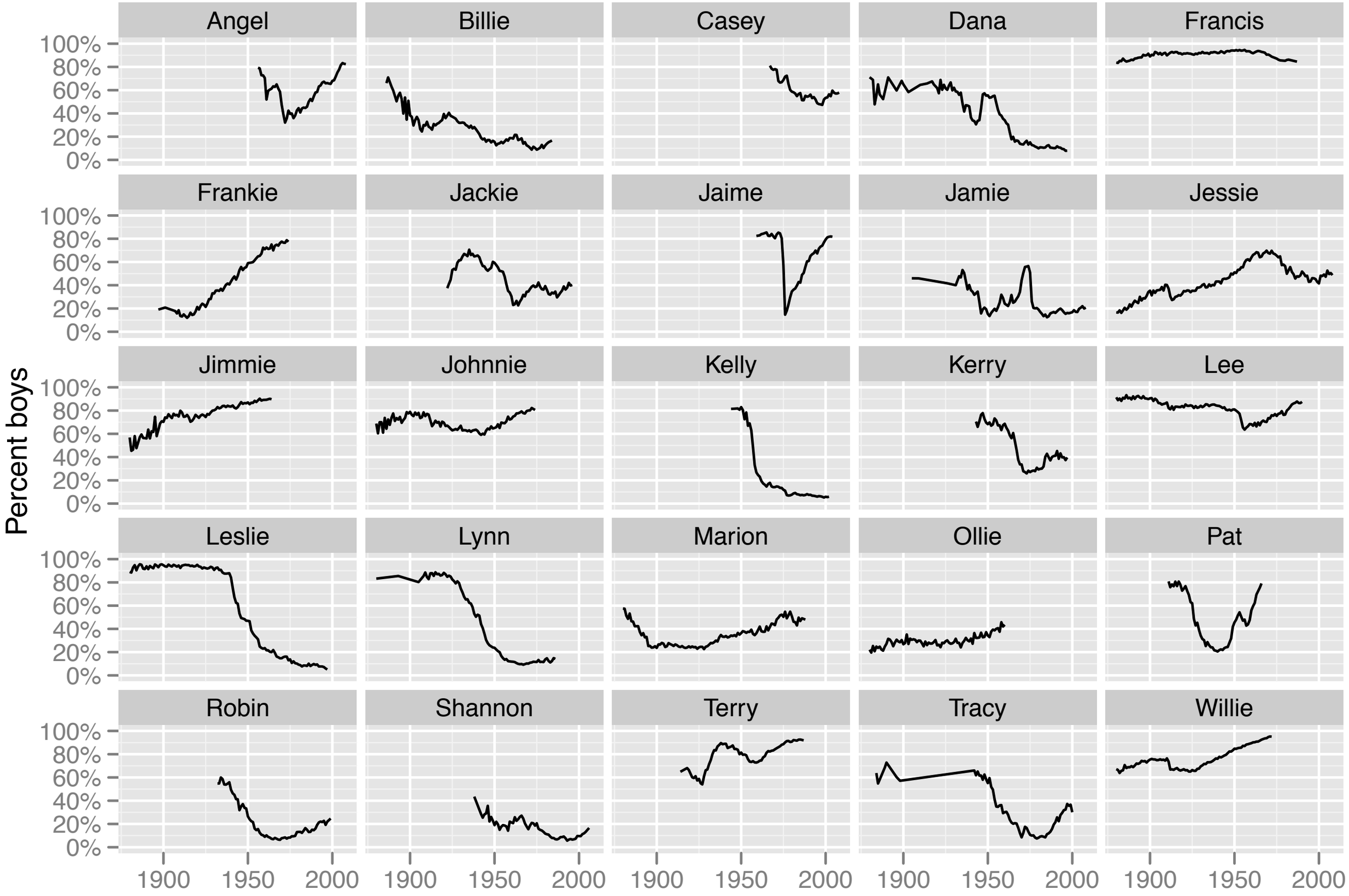
Why statistics?

“The best thing about being a statistician is that you get to play in everyone’s backyard.”

—*John W. Tukey*

“The business of the statistician is to catalyze the scientific learning process.”

—*George Box*


Angel
Billie
Casey
Dana
Francis
Frankie
Jackie
Jaime
Jamie
Jessie
Jimmie
Johnnie
Kelly
Kerry
Lee
Leslie
Lynn
Marion
Ollie
Pat
Robin
Shannon
Terry
Tracy
Willie
Percent boys
100%
80%
60%
40%
20%
0%
1900
1950
2000

Assault (homicide) by other and unspecified firearm discharge

Assault (homicide) by sharp object

Drowning and submersion while in natural water

Exposure to unspecified electric current

Motor– or nonmotor–vehicle accident, type of vehicle unspecified

Pedestrian injured in other and unspecified transport accidents

Traffic accident of specified type but victim's mode of transport unknown

Unspecified drowning and submersion

prop

0.10
0.08
0.06
0.04
0.02

5
10
15
20

hod

“I keep saying the sexy job in the next ten years will be statisticians.”

—Hal Varian
Chief Economist, Google

“The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a byproduct of the recent explosion of digital data.”

NY Times, August 09

Traditional	Recent
Pharma Biomedical Marketing Finance Government	Startup Social media Advertising Internet

Stats @ Rice

Prereqs	Core classes	Other
Math101 Math102 Math211/ Caam335/ Caam336 Math212	**Stat310** Stat405 Stat410	6 electives (2 outside stat with permission) 1 design project

Weekly lunches

Double major

Works great as a double major.

Application: biology, psychology, sociology, sports science

More tools: caam, math, cs

Minoring

Three required:
Track A: stat310, stat405, stat400/410
Track B: stat100, stat280, stat385

Three elective:
300 level+, one outside stat if it has strong statistical component

Stat410

Introduction to linear models

Powerful and general statistical tool.

Theory and data.

Offered in Fall.

Stat405

Project based introduction to data analysis. Lots of computing and hardly any maths.

http://had.co.nz/stat405

Offered in Fall (and Spring?)

Electives

SOCI 436 (Houston area survey), 313 (demography)

ECON 405 (game theory), 409 (econometrics), 475 (optimisation), 401 (math of economics), 479 (modelling)

STAT 385, 431 (more theory), 421 (time series), 422 (Bayesian data analysis), 423 (bioinformatics), 453 (biostatistics), 480 (VIGRE seminars), 485 (environmental)

Review

Inference

Sampling distributions

Sequences and limits

Bivariate distributions

Univariate continuous

Univariate discrete

Probability

Inference

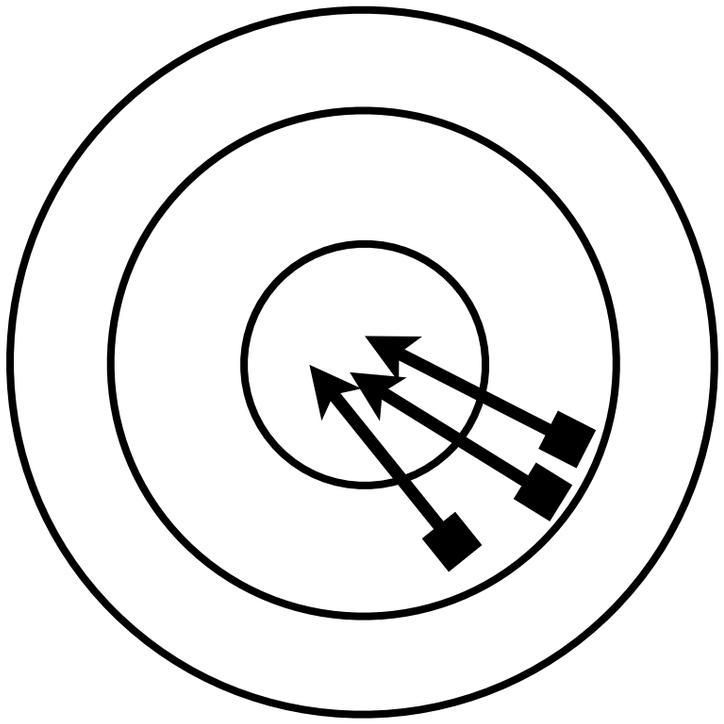
Inference

- **Point estimation**: what's our best guess for the parameter of the distribution that this data came from
- **Interval estimation**: how can we make a guess that has known probability of being right
- **Testing**: what's the probability the data really came from this distribution?

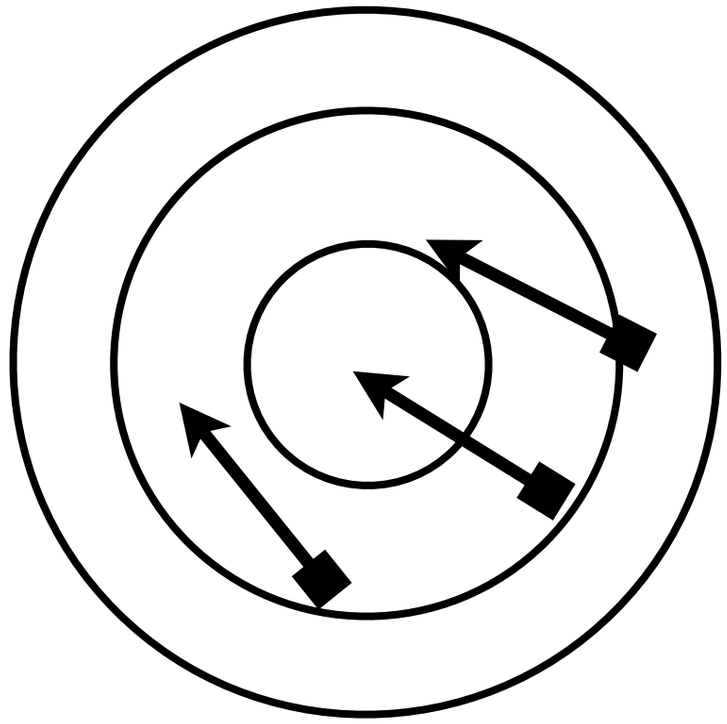
Point estimators

- **Method of moments**: easy, not always good, know nothing about it's performance
- **Maximum likelihood**: a bit harder, always consistent (often biased), is approximately normally distributed
- Properties: bias, variance, consistency

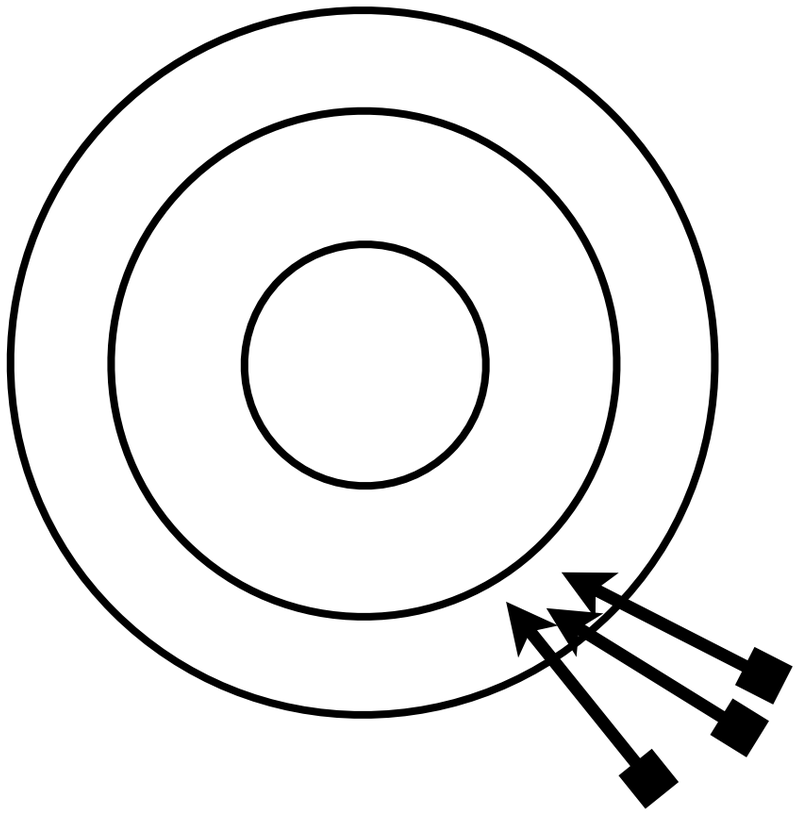
Low bias, low variance



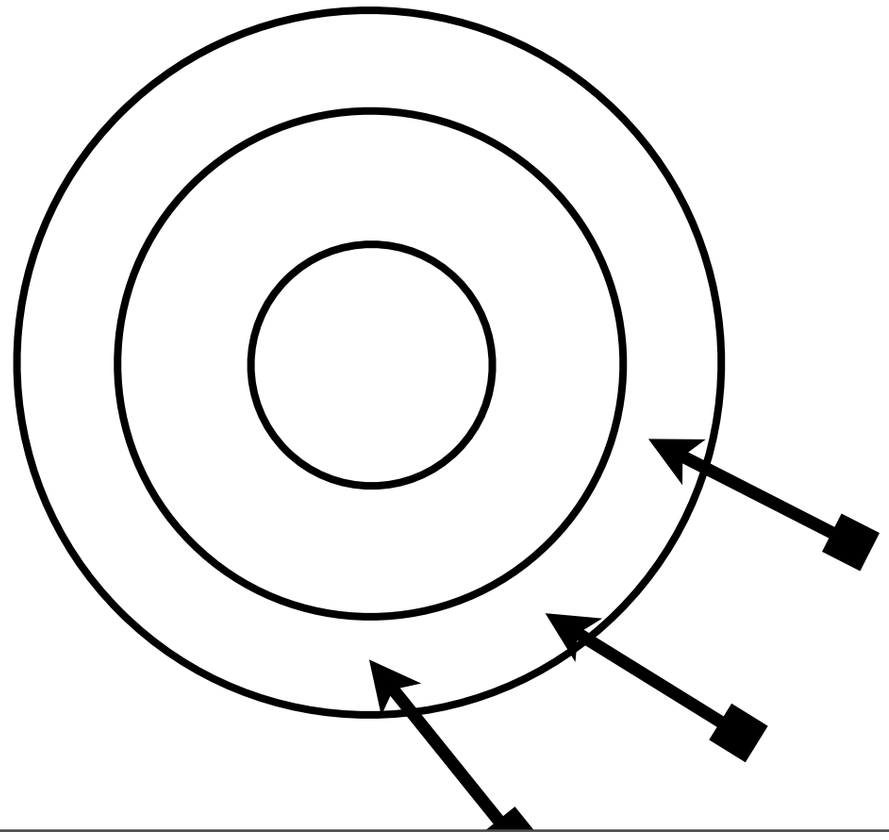
Low bias, high variance



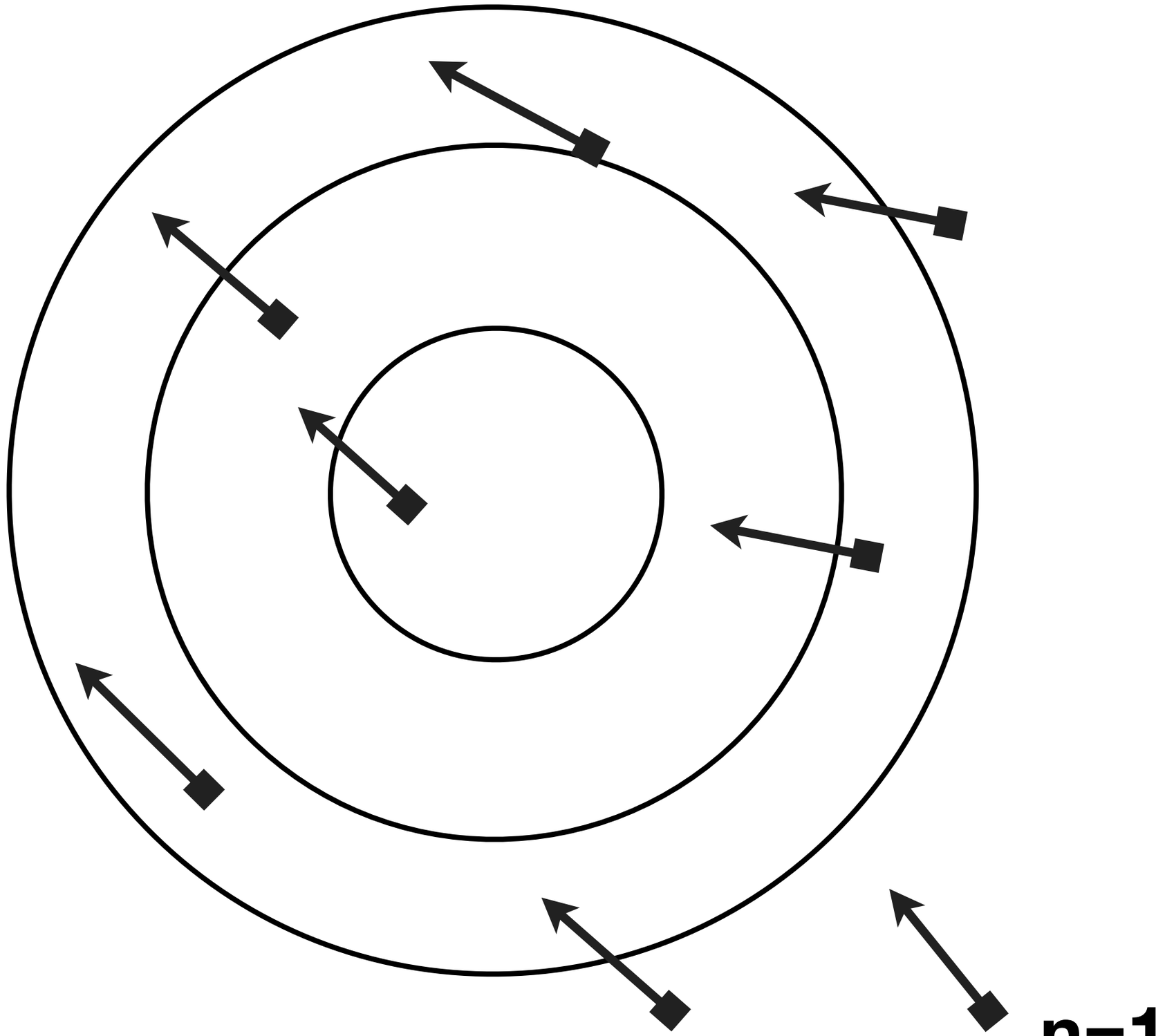
High bias, low variance



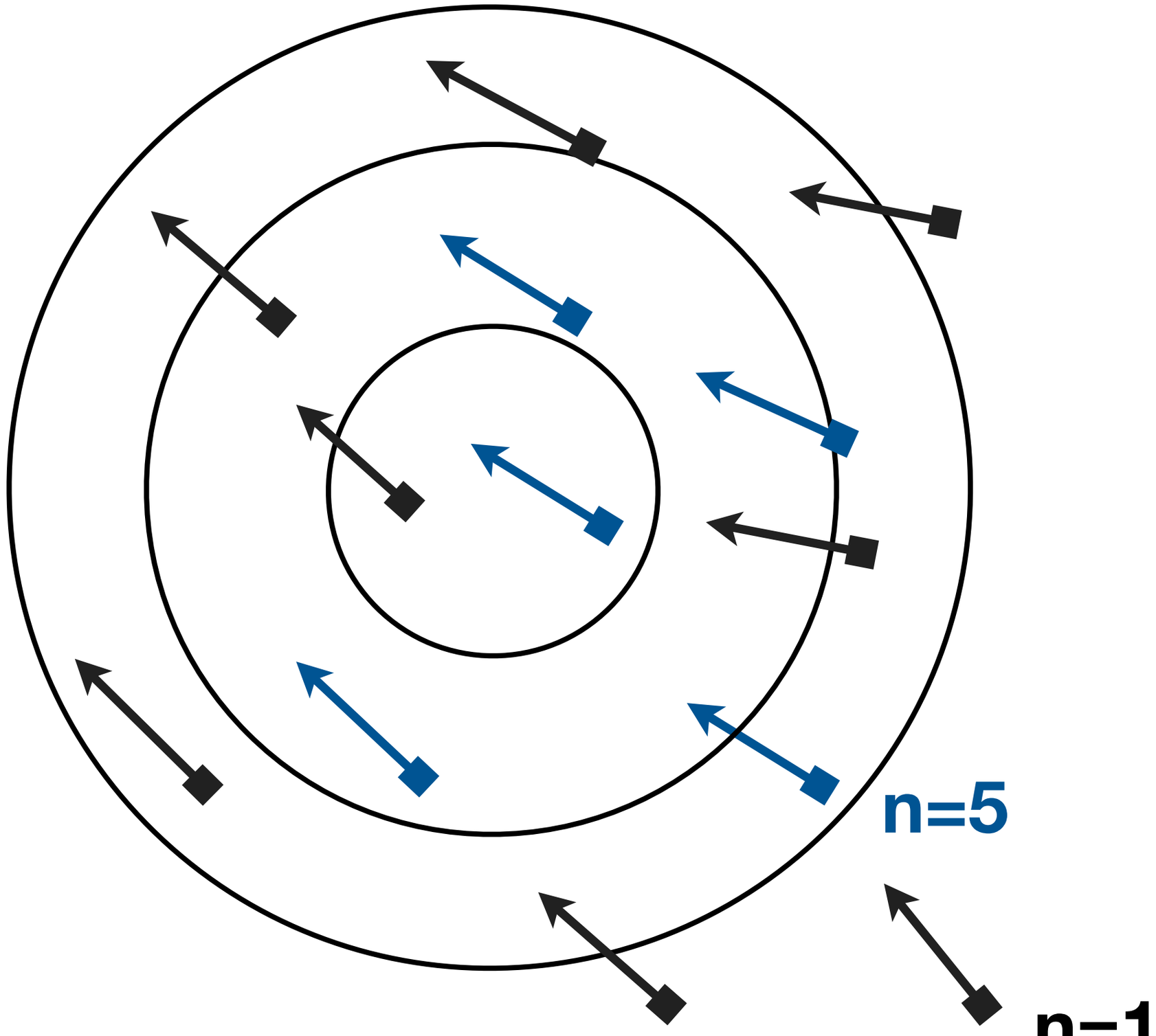
High bias, high variance



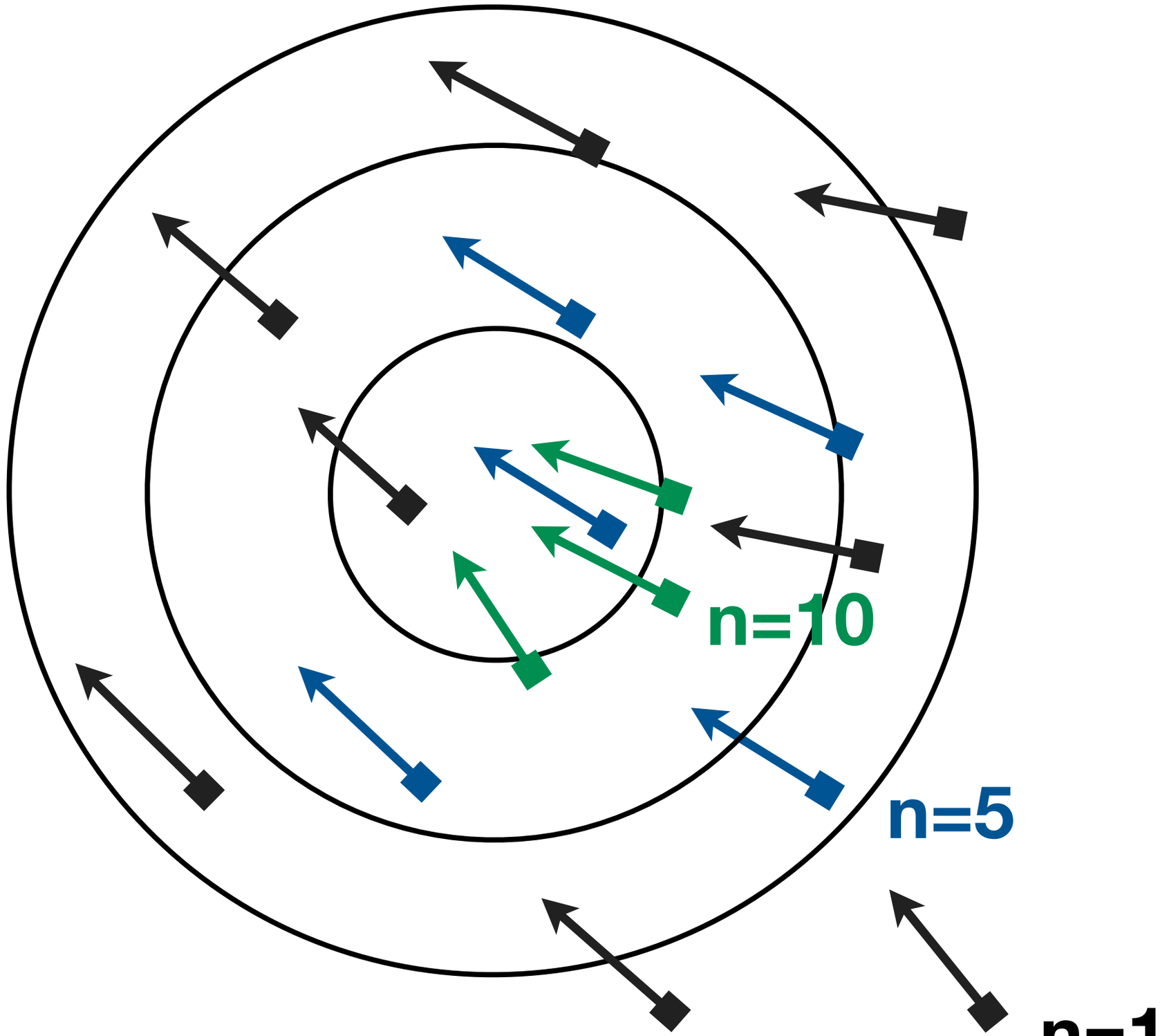
Consistent



Consistent



Consistent



Confidence intervals and testing

- Both use sampling distribution (we only know five)
- **Confidence interval**: given probability, find central region
- **Testing**: given tail region, find probability

Your turn

Compare and contrast the process of making a confidence interval with performing a hypothesis test

Confidence interval	Hypothesis test
Find distribution that connects sample to parameter	Establish Ho and Ha. Determine test statistic. Work out null distribution
Work out bounds for known dist. Back transform. Write as interval	Compute p-value. Make decision.

Sampling distributions

Your turn

What are the five sampling distributions we use most commonly?

Why do those summary statistics have those distributions?

$$X_i \overset{iid}{\sim} \text{Normal}(\mu_x, \sigma^2)$$

$$i = 1, \cdots, n$$

$$Y_j \overset{iid}{\sim} \text{Normal}(\mu_y, \sigma^2)$$

$$j = 1, \cdots, m$$

$$\bar{X}_n - \bar{Y}_m$$

Not on final!

$$T = \frac{Z}{\sqrt{V}/\sqrt{\nu}}$$

Z is normal
V is X^2 with v degrees of freedom
Z and V are independent

then

T is t-distributed with v degrees of freedom

Sequences and limits

Basic question

What is the distribution of a sum of n random variables?

What is the distribution of a mean of n random variables?

What is the distribution of some other summary of n random variables?

Your turn

What are the three useful properties of the mgf?

Why is independence so important? What property of random experiments allows us to assume it?

Important tools

MGFs

- mgf of a sum is the ...

Limits

Joint pdf is product of ... if ...

Important terms

iid

LLN

CLT

Transformations

	1d	2d
Change of variables	Must have inverse. Procedure easy	Must have inverse. Bounds and inverses can be tricky
Distribution function technique	Always works. Use definition of cdf	Easy, when it works
MGF	Only easy for sums and for known distributions. Works for any dimension	

Also remember relationship between uniform distribution and any univariate distribution

Steps

Write down u_1, u_2, $f_{X,Y}$

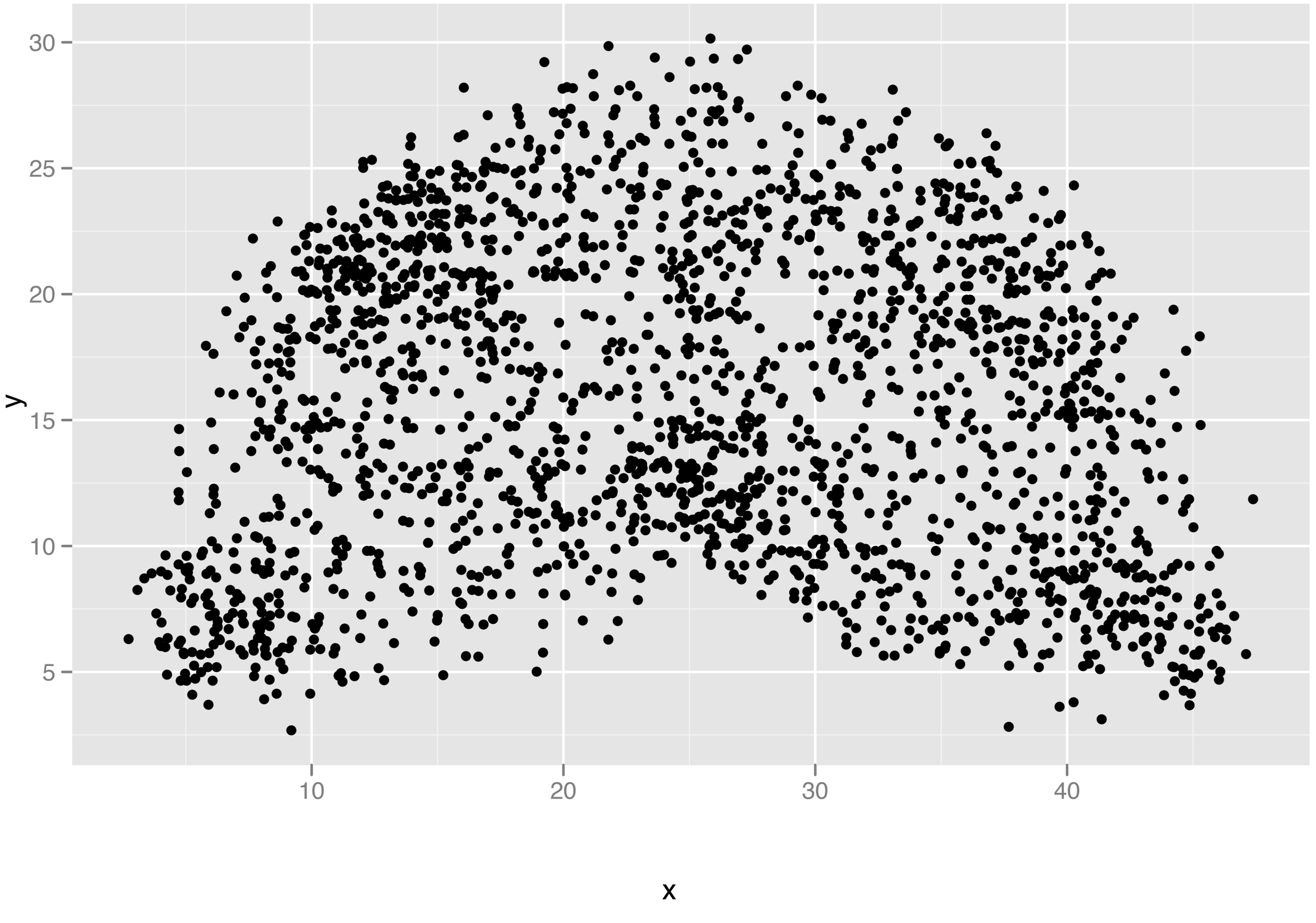
Figure out bounds of A and B

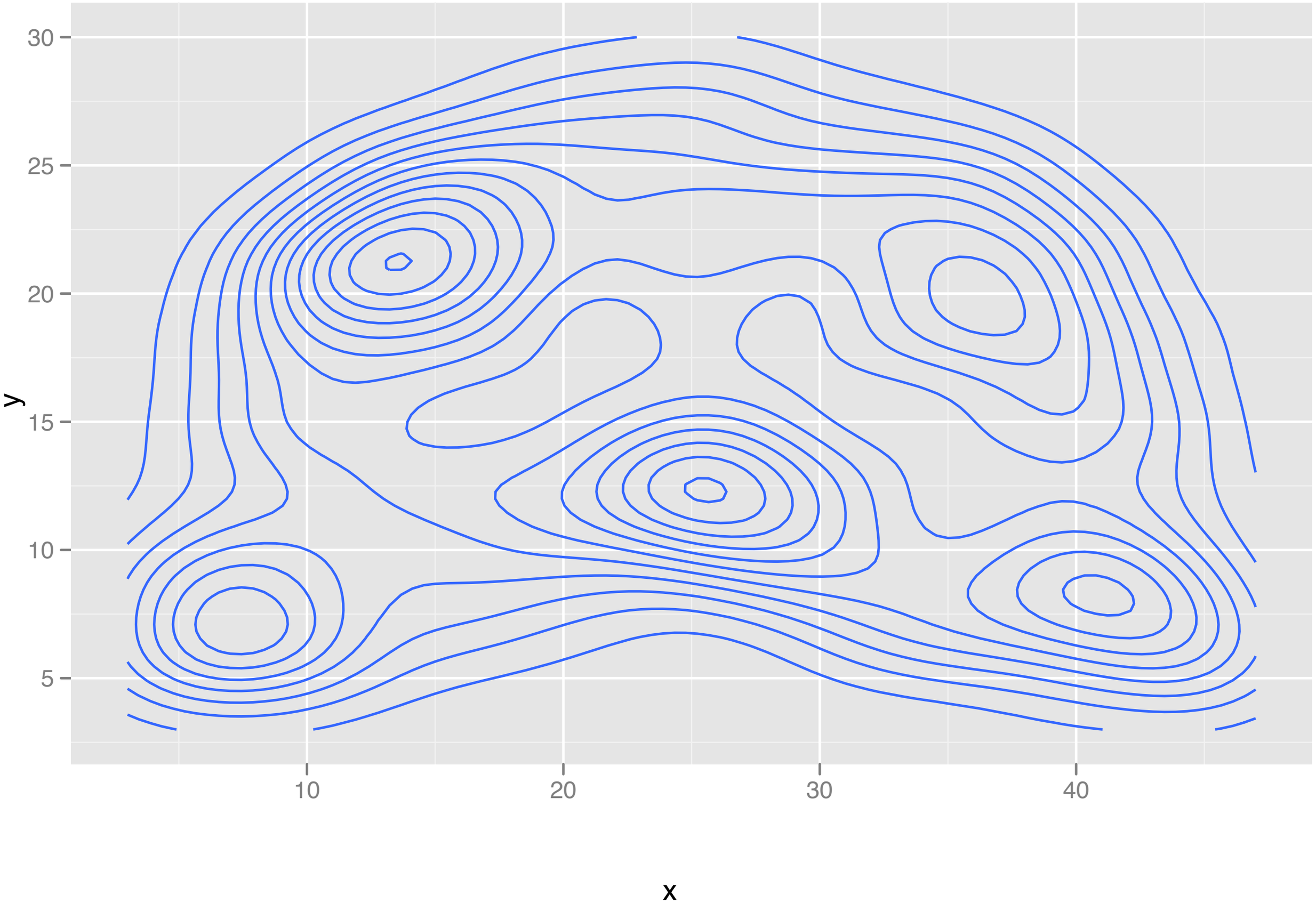
Figure out v_1 and v_2

Compute partial derivatives

Plug into formula

Bivariate random variables


30
25
20
15
10
5
y
10
20
30
40
x


30
25
20
15
10
5
y
10
20
30
40
x

2d

1d

Marginal

integration

Joint

fixing value

Conditional

$$f(x,y) = f(x|y)f(y)$$

$$f(x,y) = f(y|x)f(x)$$

X and Y are independent iff

$$f(x,y) = f(x)f(y)$$

Univariate random variables

Definitions

A **random variable** is a **random experiment** with a numeric sample space.

A **discrete** random variable has finite or countably infinite sample space. Subset of the integers. Use **sums**. Has p**m**f.

A **continuous** random variable has uncountably infinite sample space. Subset of the real line. Use **integrals**. Has p**d**f.

$$f(x)$$

For continuous x, f(x) is a probability **density** function.

Not a probability!

(may be greater than one)

Probability	Discrete	Continuous
$P(A) \geq 0$ for all $A \subset S$	$f(x) > 0$ for all $x \subset S$	$f(x) > 0$ for all $x \subset \mathbf{R}$
P(S) = 1	$\sum_{x_i \in S} f(x_i) = 1$	$\int_{\mathbb{R}} f(x) = 1$

$$E(g(X)) = \sum_{x \in S} f(x)g(x)$$

$$E(u(X)) = \int_{\mathbb{R}} u(x)f(x)dx$$

little x vs. big x

$$F(x) = P(X \leq x)$$

discrete

continuous

$$F(x) = P(X \leq x)$$

discrete $$F(x) = \sum_{t \leq x} f(t)$$

continuous

$$F(x) = P(X \leq x)$$

discrete $$F(x) = \sum_{t \leq x} f(t)$$

continuous $$F(x) = \int_{-\infty}^{x} f(t)dt$$

Outside support,
f(x) = 0

$$f(x)$$

Integrate from -∞ to x

Differentiate

$$\mathrm{F}(x)$$

Outside support,
F(x) = 0 or F(x) = 1

Moments

The ith **moment** of a random variable is defined as $E(X^i) = \mu'_i$. The ith **central moment** is defined as $E[(X - E(X))^i] = \mu_i$

The mean is the _____ moment. The variance is the _____ central moment.

Your turn

List all the named distributions we know about. When do you use each one?

Distribution	Special property
Binomial	Adding two binomials with same p is binomial
Poisson	Waiting t times as long is also Poisson
Exponential	Memoryless. Waiting time for between Poisson events.
Gamma	Adding two gammas with same rate/wait is still gamma
Normal	Adding two normals is a normal. * or + constant is still normal.

Probability

Random experiment

"A **random experiment** is an experiment, trial, or observation that can be repeated numerous times under the same conditions... It must in no way be affected by any previous outcome and cannot be predicted with certainty." (http://cnx.org/content/m13470/latest/)

i.e. it is **uncertain** (we don't know ahead of time what the answer will be) and **repeatable** (ideally).

Sample space

The **sample space** is the **set** containing all possible **outcomes** from a random experiment. Often called S.

In set theory called **U**
For rv's, often called the **support**

More terminology

An **event** is a ...

A collection of events are **mutually exclusive** if...

A collection of events are **exhaustive** if...

A collection of events is a **partition** if ...

Regular	Conditional	
$P(A) \geq 0$	$P(A \mid C) \geq 0,$	for all $A \subset S$
$P(S) = 1$	$P(S \mid C) = 1$	
$P(A \cup B) = P(A) + P(B)$	$P(A \cup B \mid C) = P(A \mid C) + P(B \mid C)$	if $A \cap B = 0$

Regular	Conditional
$P(A') = 1 - P(A)$	$P(A' \mid C) = 1 - P(A \mid C)$
$P(A) \leq P(B)$ if $A \subset B$	$P(A \mid C) \leq P(B \mid C)$ if $A \subset B$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	$P(A \cup B \mid C) = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$

Independence

If A and B are **independent**, then

$P(A \cap B) = ...$

This implies $P(A \mid B) = ... ?$

Events are **mutually independent** if ...

Toolbox

Complements

Convert union to sum

Convert intersection to conditioning
(and vice versa)

Convert intersection to product
(if independent)

Use law of total probability

Switch conditioning (Bayes' rule)

Counting

- Multiplication principle
- If order doesn’t matter, divide by total possibilities by number of ways of reordering them

Feedback

Feedback

Did you like the study skills advice? More or less?

How could I make the homework help sessions better? How were the homeworks? Too long? Too short? Too many? Too few?

How was the assessment overall? Too much emphasis on homeworks? Too much on exams?