Week 2: Review of probability and statistics

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Outline

- The very basics of stats
- Random variables; joint distributions
- Independence, conditional distributions, conditional independence
- Expected value, conditional expectation, variance
- A model to understand statistical inference: population and samples
- Properties of an estimator: bias, efficiency, and consistency

Big picture

- Carefully review Wooldridge Appendix A to C
- Dust off your calculus notes for a refresher or see videos in Concepts to Know
- Do NOT focus on the mechanics. Do remember how to take basic derivatives
- Focus on **concepts**

Basics

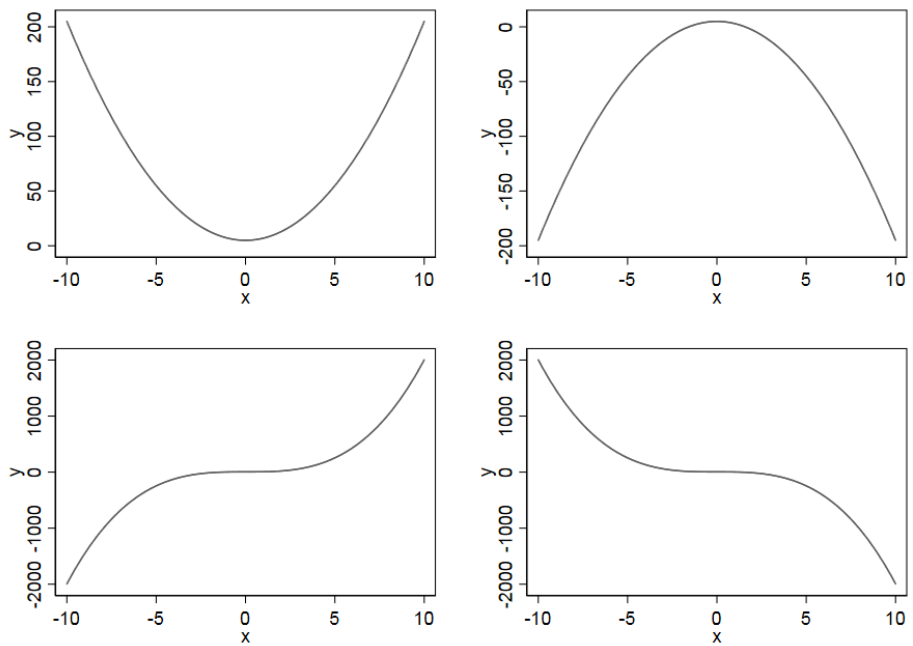
- Rules of summation and averages
  1. $\sum_{i=1}^{n} c = nc$
  2. $\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i$
  3. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ ("demeaning")
  4. $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$
  5. And, of course: $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Recall what is a linear function: $y = a + bx$
- In a linear function: $\Delta y = b\Delta x$
- We use calculus for small changes: $\frac{dy}{dx} = b$

Other functions

- Quadratic: $y = a + bx + cx^2$
- Cubic: $y = a + bx^3$
- Polynomials: $y = a + bx + cx^2 + dx^3 + ex^4$
- Log function: $y = log(x)$
- Exponential: $y = exp(x)$
- **Remember how to take the derivative of these functions**

As usual, Stata is your friend – learn to use Mathematica, too

```
twoway function y= 5 + 2*x^2, range(-10 10) saving(gf1.gph, replace)
twoway function y= 5 - 2*x^2, range(-10 10) saving(gf2.gph, replace)
twoway function y= 5 + 2*x^3, range(-10 10) saving(gf3.gph, replace)
twoway function y= 5 - 2*x^3, range(-10 10) saving(gf4.gph, replace)
graph combine gf1.gph gf2.gph gf3.gph gf4.gph
graph export gf_all.png, replace
```



Derivatives are easy to understand conceptually

- The rules for derivatives come from the **definition** of the derivative

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

- Limits are central to calculus
- So, **calculating the derivative of any function reduces to taking the above limit**. That's it!
- For example, the derivative of $f(x) = log(x)$ is $f'(x) = \frac{1}{x}$
- So the slope of $log(x)$ is not constant (linear). At, say, $x = 10$, it's 0.1. Note another thing. The larger x gets the smaller the slope gets. Or in calculus jargon, the limit of $f'(x)$ when x tends to infinity is zero. If x is small, the slope is large

Remember the basics

- The rules for derivatives are found by taking the limit of h tending to zero:

Proof We start, of course, with the difference quotient that defines the derivative, and we then simplify it using the basic properties of the logarithm. Fix $x > 0$.

$$\frac{\ln(x+h) - \ln x}{h} = \frac{1}{h} \ln\left(\frac{x+h}{x}\right) = \ln\left(1 + \frac{h}{x}\right)^{\frac{1}{h}}$$
$$= \ln\left(1 + \frac{1/x}{1/h}\right)^{\frac{1}{h}}.$$

Now, let $m = 1/h$. As $h \to 0$, $m \to \infty$. Continuing our calculation with $m = 1/h$, we find

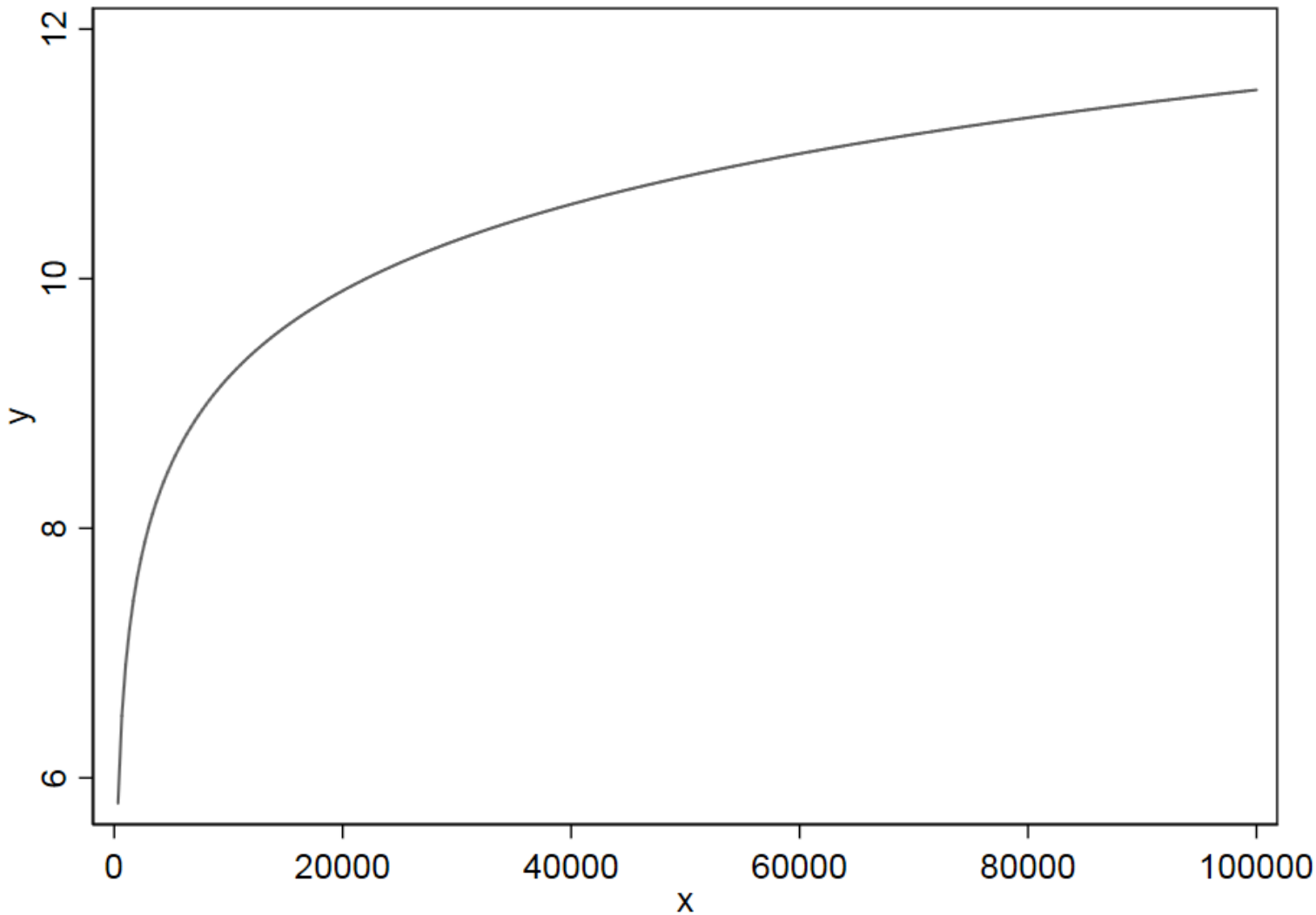
$$\lim_{h\to 0} \frac{\ln(x+h) - \ln x}{h} = \lim_{m\to\infty} \ln\left(1 + \frac{1/x}{m}\right)^{m}$$
$$= \ln \lim_{m\to\infty} \left(1 + \frac{1/x}{m}\right)^{m}$$
$$= \ln e^{1/x} = \frac{1}{x}.$$

Therefore, $(\ln x)' = 1/x$. The fact that we can interchange ln and lim in the above string of equalities follows from the fact that $y = \ln x$ is a continuous function: $x_m \to x_0$ implies that $\ln x_m \to \ln x_0$; or equivalently,

$$\lim_m (\ln x_m) = \ln\left(\lim_m x_m\right). \quad \blacksquare$$

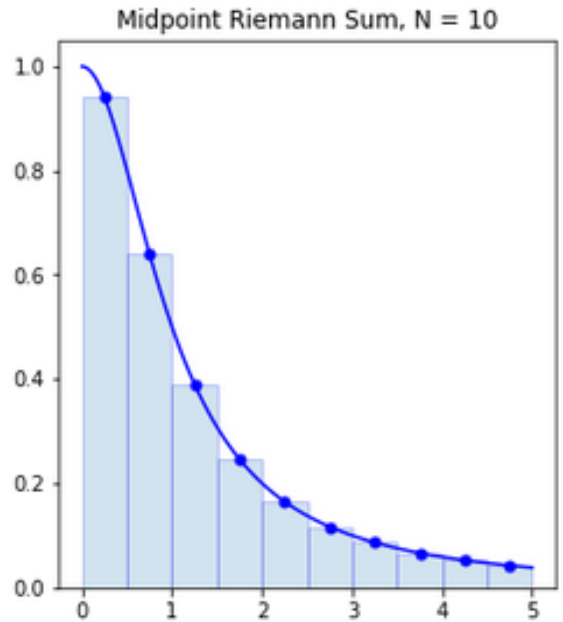
As I said, Stata is your BFF

```
twoway function y= log(x), range(-5 100000) saving(gf5.gph, replace)
graph export gf_log.png, replace
```



What about integrals?

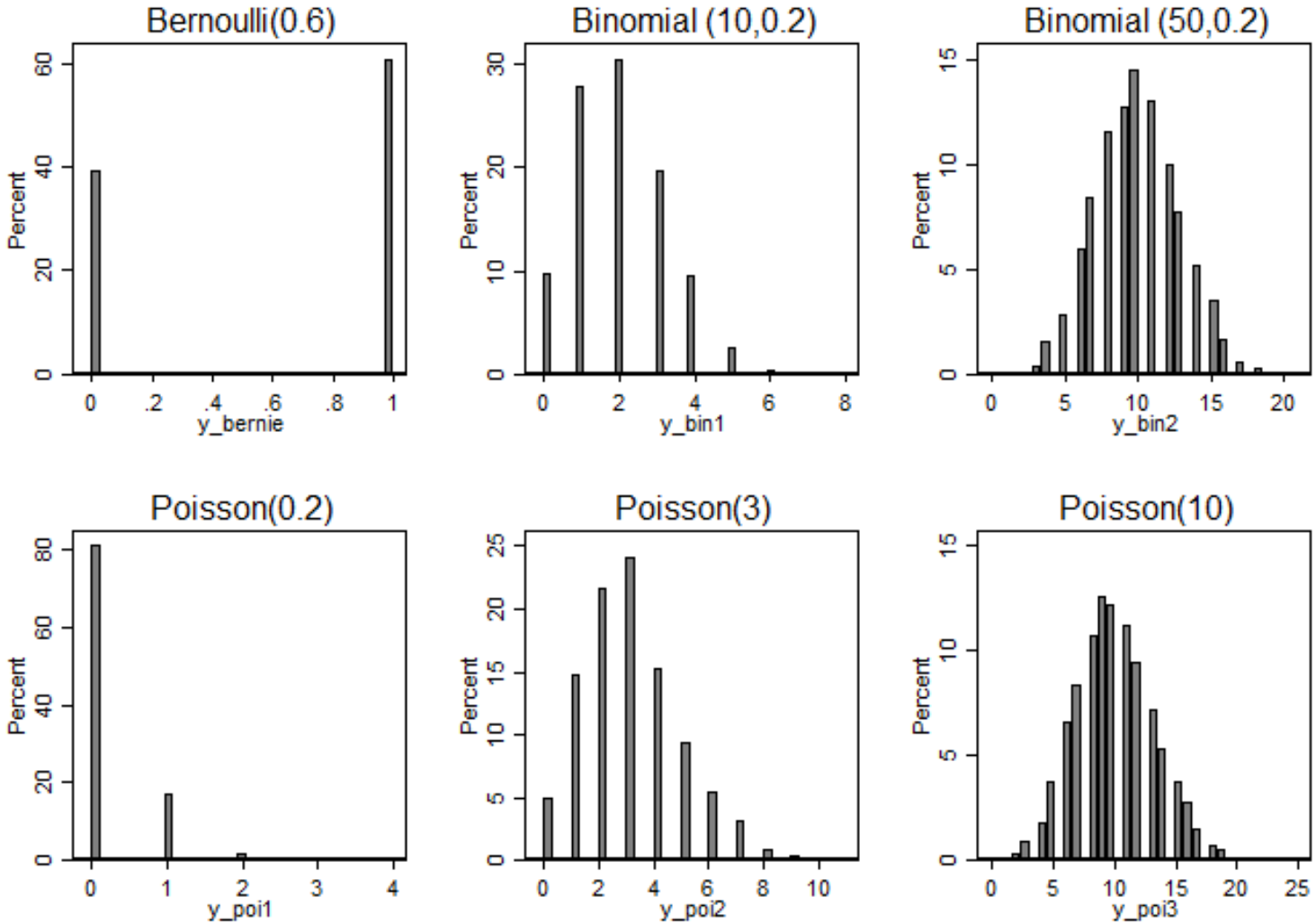
- You learn sometimes that they are the opposite of the the derivative
- We care here about the **definitive integral**, with is the area under a curve
- Break the area under a curve into rectangles and then make the rectangles small
- We use Reiman sums to approximate, summing over the area of the small rectangles



Random variables

- A **discrete** random variable takes only a finite or **countably** infinite number of values
- We *describe* a random variable by listing the possible values it takes and the probability of taking each of those values; probabilities must add up to 1
- The **probability density function** (pdf) is a *summary* of all the possible **outcomes** and their **probabilities**
- The simplest, Bernoulli: $f(x; p) = p^x(1-p)^{(1-x)}$, where $x \in \{0, 1\}$
- p is a **parameter**, which describes the probability of "success" or 1
- The **support**, x, is the values a random variable can take: only 0, 1
- Binomial has two parameters, n and p: $f(k; n, p) = \binom{n}{k} p^k (1-p)^{(n-k)}$

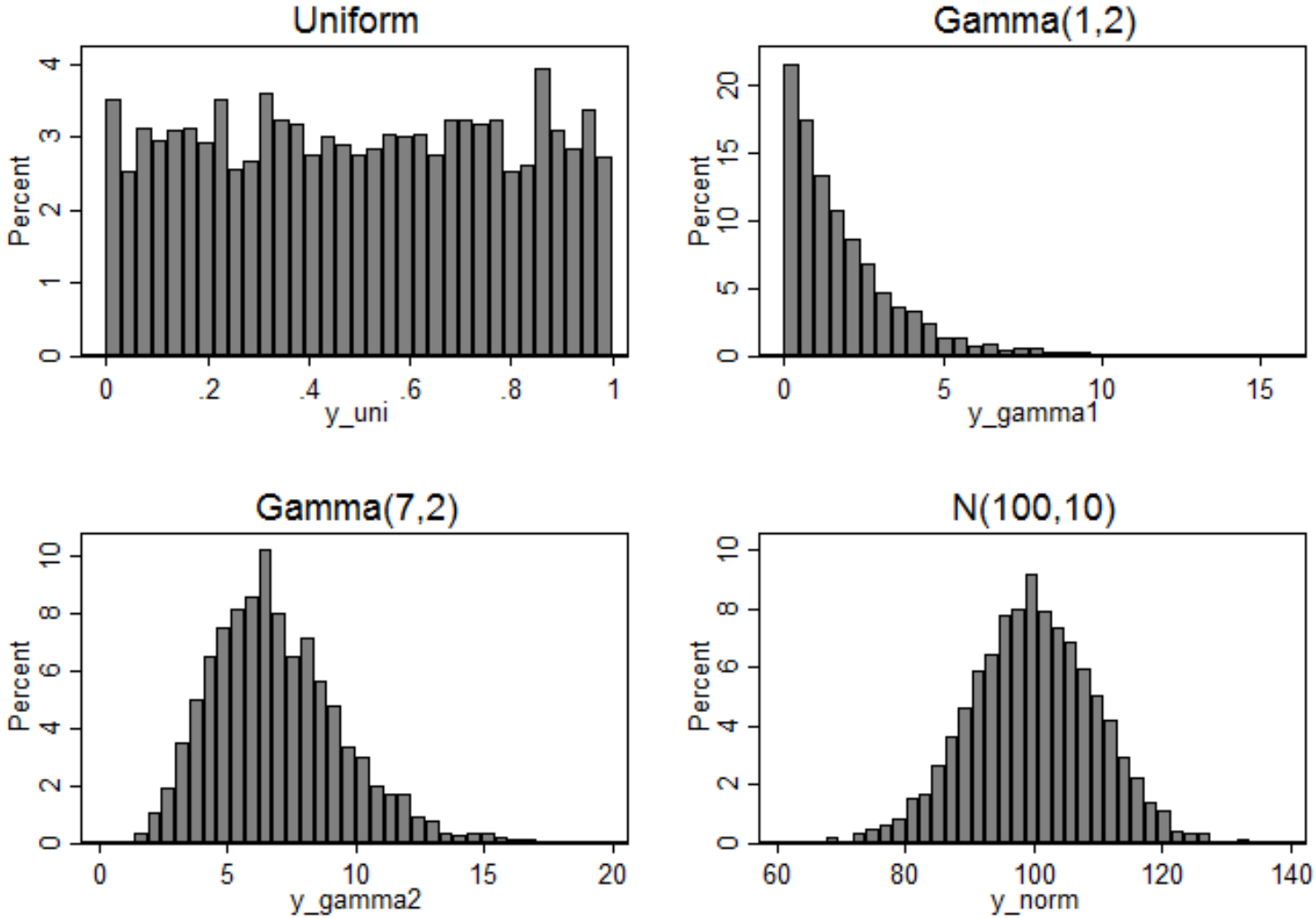
Examples of discrete distributions



Continuous random variables

- A **continuous** random variable can take infinitely may values
- The **probability of taking a single value is zero** (not exactly intuitive)
- So we can only calculate the probability of a **range** of values: $P(a \leq X \geq b)$ or $P(a < X > b)$
- The most common example, **normal distribution**. The density function is: $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- To calculate the probability we need to integrate: $P(a \leq X \geq b) = \int_a^b f(x; \mu, \sigma)$
- Because probabilities must add up to 1: $\int_{-\infty}^{+\infty} f(x; \mu, \sigma) dx = 1$
- **Cumulative distribution function** (cdf): $F(x) = P(X \leq x)$; for normal: $\int_{-\infty}^{x} f(x; \mu, \sigma)$

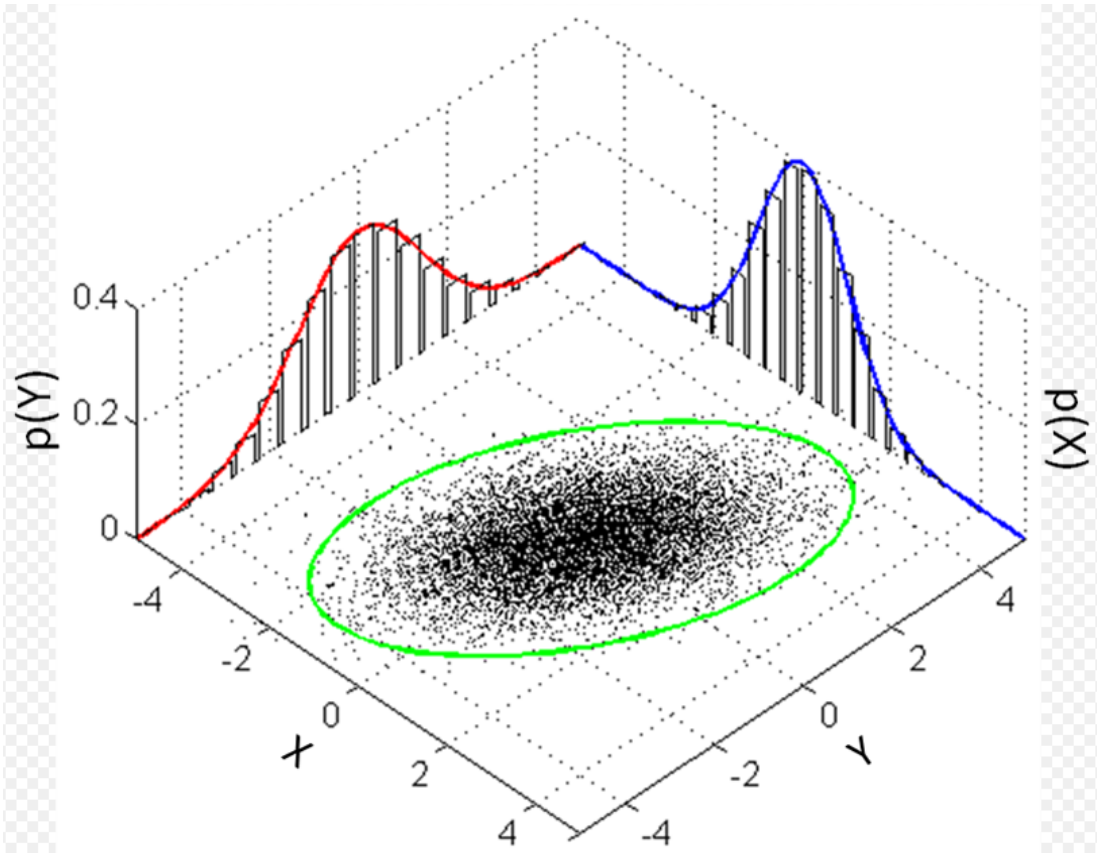
Examples of continuous distributions



Joint distributions

- We often care about **more than one random variable** at a time
- It's straightforward to extend the notion of probability density function to a **joint probability density** function:
- For a discrete rv: $f_{X,Y}(x, y) = P(X = x; Y = y)$
- Needs to add up to 1: $\int_x \int_y f(x, y) dx dy = 1$, for continuous case
- We won't directly deal with joint distributions this semester although some **key statistical concepts** emerge from joint distributions

Bivariate normal: $N(\mu, \Sigma)$



- Graph from Wikipedia
- Note that to simulate, we need to specify the correlation besides the usual mean and sd for each normal

Marginal distributions

- Remember the concept of marginal distributions (**absolutely NOTHING to do with marginal effects**, by the way)
- In the case of two random variables (X,Y), the marginal distribution of, say, X is the probability of X averaging over all the information on Y. The other way of understand it, the marginal distribution of X gives you the probability of each of the values X can take regardless of the values of Y (it has to add up to 1).
- From Wikipedia:

Y \ X	x_1	x_2	x_3	x_4	$p_Y(Y)\downarrow$
y_1	4/32	2/32	1/32	1/32	**8/32**
y_2	2/32	4/32	1/32	1/32	**8/32**
y_3	2/32	2/32	2/32	2/32	**8/32**
y_4	8/32	0	0	0	**8/32**
$p_X(X)\rightarrow$	**16/32**	**8/32**	**4/32**	**4/32**	**32/32**

Joint and marginal distributions of a pair of discrete random variables X, Y having nonzero mutual information $I(X; Y)$. The values of the joint distribution are in the 4×4 square, and the values of the marginal distributions are along the right and bottom margins.

Independence

- Two random variables are independent if and only if: $f_{X,Y}(x, y) = f_X(x)f_Y(y)$
- For discrete rv: $P(X = x, Y = y) = P(X = x)P(Y = y)$
- In terms of events: $P(A \cap B) = P(A)P(B)$
- This definition of independence is not that intuitive
- It helps if you think that you **don't need the joint** density to figure out the probability that two events happen; you just need the probability of each event or the **marginal probabilities** ($f_X(x)$ and $f_Y(y)$)
- Independence means that knowing the outcome of one variable does not affect the probability of the other
- Easier to understand once we define conditional distributions

Conditional distributions

- What is the probability that a (discrete) random variable Y takes the value of y **given** that $X = x$?
- Think of "given" as meaning **when**, **if**, **assuming that**... $X = x$
- When you think about events: What is the probability that event A happens given than B already happened?: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Continuous joint distributions: $f_{Y|X} = \frac{f_{X,Y}(x,y)}{f_X(x)}$
- Another way of thinking about conditional distribution: **the probability space is being restricted**

Independence, again

- What is the conditional probability if two events are independent?
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$
- So the probability of of A given than B occurs is just $P(A)$
- In words, B happening does not affect $P(A)$ (and vice versa)
- The notion of **given** or **conditional** will come back many, many times during the semester in many forms
- **Note**: We're talking about probabilities here, but a related concept is **conditional expectation**. Example: What is the average salary for a professor in the US? What is average salary conditional on rank? What is the salary given than the professor is male?

Conditional independence

- Conditional independence is a key related concept. What about if the relationship between A and B is affected by another factor, Z?
- A and B are **conditionally independent** if $P(A \cap B|Z) = P(A|Z)P(B|Z)$
- More useful: If A and B are conditional independent given Z: $P(A|B, Z) = P(A|Z)$
- In words, knowing B doesn't tells us anything about $P(A)$ **once we know Z**

Conditional independence...

- Usual example: probability of two people commuting back home from work (third factor: location)
- Another example: Supposed you randomly assign patients to an experimental treatment T based on severity of illness I. People with more severe illness are given more chances of receiving the treatment. Furthermore, age A is positively related to severity of illness
- Therefore, T and A are **NOT independent** but T and A are **conditionally independent** given I:
- $P(T|A, I) = P(T|I)$

Expected value

- The **expected value** of a random variable is the average of all the values that X takes **weighted** by their respective probabilities
- Discrete rv: $E(X) = x_1 f(x_1) + \cdots + x_k f(x_k) = \sum_{i=1}^{k} x_i f(x_i)$
- Continuous rv: $E(X) = \int_{-\infty}^{+\infty} x f(x) dx$
- Note that the mean is an expected value with equal weights: $\sum_{i=1}^{n} \frac{x_i}{n}$
- Properties
  1. $E(c) = c$ for any constant c
  2. $E(cX) = cE(X)$
  3. $E(aX + b) = aE(X) + b$ for constants a, b
  4. $E(X + Y) = E(X) + E(Y)$

Conditional expectation

- What is the expected value of a random variable Y given that we know $X = x$ has occurred?
- Discrete case: $E(Y|X = x) = \sum_{i=1}^{m} y_i f_{Y|X}(y_i|x)$
- Continuous case: $E(Y|X) = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x)$
- It looks more complicated that what it is: What is the average temperature in Denver over the year? What is the average temperature during winter?
- The **law of iterated expectations**: $E(E(Y|X)) = E(Y)$
- Obvious but not intuitive: We first calculate the expected value $E[Y|X]$. Say, the average weather by season. It's a function of the four seasons. So it's a function of X, where X is season
- We then calculate $E(E(Y|X))$ and get $E(Y)$. Same is taking the average weather for all four seasons...

Another way of understanding the law of iterated expectations

- Suppose that $E(wage|educ) = 4 + 0.6 * educ$
- That is, we have a linear functional form for wage given levels of education (educ)
- The law of iterated expectations tell us that we can obtain $E(wage)$ by replacing $educ$ with $E(educ)$
- $E(wage) = 4 + 0.6 * E(educ)$
- In this class, we will learn how to estimate something like $E(wage|educ) = 4 + 0.6 * educ$ from data

Variance

- The variance is a measure of how close observations are to their mean
- $Var(X) = E[(X - \mu)^2] = E(X^2) - (E(x))^2 = E(X^2) - \mu^2$
- Properties
  1. The variance of a constant is zero
  2. $Var(aX) = a^2 Var(X)$
  3. $Var(aX + b) = a^2 Var(X)$
- The standard deviation is $sd(X) = +\sqrt{Var(X)}$
- The variance is a measure of deviation from the "mass" of the distribution
- The **normal distribution is unique in that the variance is NOT a function of the mean**

Populations and samples

- Statistical inference is based on the idea that there is a population (for technical reasons, **infinitely large**)
- We want to make inferences about the population but we can't measure the infinitely large population; instead, we sample from it
- **Statistics is about making inferences** ("conclusion reached on the basis of evidence and reasoning") **about the population from samples**
- Not always very clear... What is the population? Do we need statistical inference?
  1. How are people going to vote in the next US election?
  2. What is the proportion of for-profit nursing homes that are certified by CMS?
  3. How many stars of type F are there in the universe?

A note on notation

- Textbooks distinguish population from sample statistics/parameters: μ vs $\bar{x}$ or σ^2 vs s^2
- Not always consistent in notation
- $\bar{x}$ and s^2 are **estimators** (from sample) of the true but unknown **parameters** μ and σ^2
- This semester we're going to distinguish β_j from $\hat{\beta}_j$
- For example, we will say that the population model is $y = \beta_0 + \beta_1 x + \epsilon$ and the estimated model from a sample is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Sampling

- Let Y be a **random variable** representing a population with **probability density function** (pdf) $f(y; \theta_1, \theta_2)$, which depends on two parameters, θ_1 and θ_2
- A random sample of size n from Y is denoted by $Y_1, Y_2, ..., Y_n$
- An actual sample is denoted by $y_1, y_2, ..., y_n$
- We say that a random sample from $f(y; \theta_1, \theta_2)$ Y_i are **independent, identically distributed (i.i.d)**
- **Important**: Note that **each** y_i is a draw from a random variable
- We will focus on iid observations in this semester; **cross-sectional** data. Longitudinal or clustered data are not iid
- Understanding statistics is a lot easier with simulations

Example 1: Normal

- Let's sample from a normal distribution with $\mu = 50$ and $\sigma^2 = 81$

```
set seed 1234567
set obs 2500
gen y = rnormal(50, 9)
sum y
. sum y

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |      2,500    50.05014    8.946829   17.52741   93.08252
list y in 1/10, noobs clean
41.41905 54.61874 65.79002 52.44608 52.50088 56.86109 45.38539
45.94027 28.07498 38.66535
sum y if _n <=10
   Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |         10    48.17018    10.63055   28.07498   65.79002
```

- **Note**: If I gave you a list of numbers and told you that they came from a normal distribution but ask you to guess μ and σ^2 and you give me the **optimal/best answer**, you would have discovered **maximum likelihood estimation** (MLE)

Example 1: Bernoulli

- Sample from *Bernoulli*(θ), where $\theta = 0.6$ (the probability of "success")

```
gen y_bernie=uniform()<=0.60
sum y_bernie
   Variable |        Obs        Mean    Std. Dev.       Min        Max
------------+---------------------------------------------------------
    y_bernie |      2,500       .5932    .4913352          0          1
list y_bernie in 1/10, noobs clean
0 0 0 1 1 1 0 1 0 0
sum y_bernie if _n <=10
  Variable |        Obs        Mean    Std. Dev.       Min        Max
------------+---------------------------------------------------------
    y_bernie |         10          .4    .5163978          0          1
```

- **Note that the mean is the proportion of 1's**. This is how we can simulate a logistic model

Properties of an estimator

- We have a model of how things work: we have a population, a sample, and we come up with an estimator of a population parameter (e.g. mean, proportion, variance, etc)
- The next question is: what is a **good** estimator? What are the **properties** of a good estimator?
- We just saw that if we take a sample from a population and calculate the mean, $\bar{x}$, the sample mean is not going to be the same as μ but it could be close. It's a sample after all
- We could take more samples and calculate the mean again of each sample. At the very least, we would like, on average, $\bar{x}$ to be the same as μ
- An estimator is **unbiased** if $E(\bar{x}) = \mu$; more generally, $E(\hat{\theta}) = \theta$

Example

- We will create a **population** of $N = 50,000$

```
clear
set obs 50000
gen y = rgamma(2,4)
* "true" mean and sd:
sum y
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           y |     50,000    8.009844    5.679751    .010425   54.47659
```

- We will then get 100 random samples and calculate the mean for each of the 100 samples

```
postfile buffer meanhat using sampmeans, replace
forvalues i=1/100 {
   preserve
   sample 30, count
   qui sum y
   post buffer (r(mean))
   restore
}
postclose buffer
```

Example...

- Let's take the mean of the 100 means

```
sum meanhat
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     meanhat |        100    7.939838    .9698508   5.713463   10.68582
```

- It's close to 8.009. **Of course, this is not a proof** but the analytic proof is very simple
- We'd need to show that the expected value of a Gamma rv is $\alpha\beta$, where α and β are the scale and shape parameters of the Gamma distribution
- What is the the Std. Dev. .9698508 called?

Example...

- We got 100 simulated means. The mean of the 100 means is 7.93. What about the 95% interval?

```
* Calculate percentiles of the 100 means
use sampmeans, clear
egen p2_5 = pctile(meanhat), p(2.5)
egen p97_5 = pctile(meanhat), p(97.5)
sum p*
  Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       p2_5 |        100    6.191483           0   6.191483   6.191483
      p97_5 |        100    9.849985           0   9.849985   9.849985
```

- If we **repeated the experiment** many times (100 here), then 95% percent of the time the mean would be between 6.19 and 9.8
- That's a **frequentist confidence interval**

Example...

- In real life examples, we do not repeat sampling. We rely on statistical theory to derive properties of estimators based on sample values. But theory relies on this notion of repeating things
- In real life, we would get one sample and use theory to tell us how to calculate the 95% CI and standard error:

```
set obs 50000
gen y = rgamma(2,4)
* get one sample
preserve
   sample 30, count
   mean y
restore
------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------
           y |   8.478952   1.001921       6.429792    10.52811
------------------------------------------------------------
```

- In your homework, you will learn (or, hopefully, remember) about the central limit theorem...

Example of a biased estimator

- The sample variance $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$ is **biased**
- $E[s^2] = E[\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}] = \frac{1}{n} E[\sum_{i=1}^{n} (x_i - \bar{x})^2]$

Proof.

$$
\begin{aligned}
E[\sum_{i=1}^{n} (x_i - \bar{x})^2] &= E[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2] \\
&= n(\mu^2 + \sigma^2) - n(\mu^2 + \frac{\sigma^2}{n}) \\
&= (n-1)\sigma^2
\end{aligned}
$$

□

- So $E[s^2] = \frac{(n-1)\sigma^2}{n}$
- Since we want $E[s^2] = \sigma^2$, we divide by $(n-1)$, not by n

Efficiency

- If we have two **unbiased** estimators, say, e_1 and e_2 of a parameter θ, we prefer the estimator that has the **smallest variance**. We call it the relatively **efficient** estimator
- Another way: If $var(e_1) \leq var(e_2)$ then e_1 is a relatively efficient estimator
- A classic example: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an efficient estimator of the population mean μ compared to y_n, the last observation of a sample
- Both are unbiased estimators but the variance of $\bar{x}$ is $\frac{\sigma^2}{n}$ while the variance of y_n is σ^2
- Easy to prove analytically

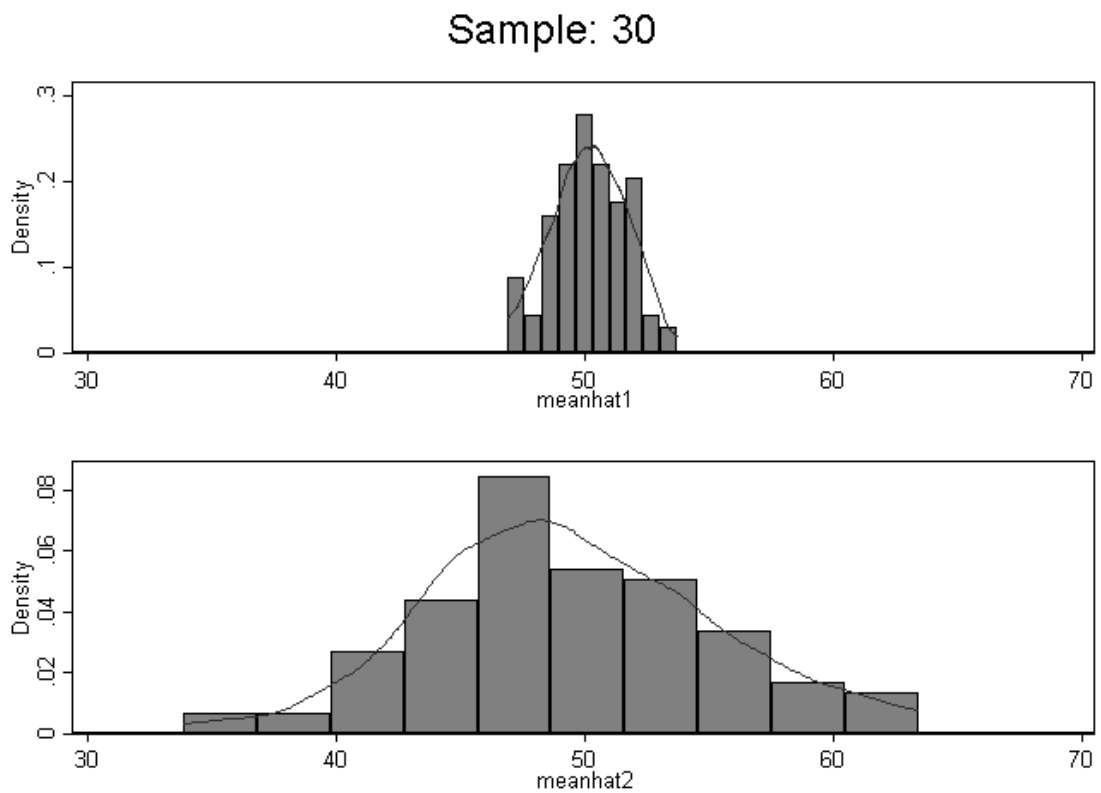
Efficiency

- We can also simulate. We will compare two estimators. The usual mean and the mean of the first and the last observation (note the syntax y[1] and y[30] in Stata). Both are **unbiased** estimators (why?)

```
set obs 50000
gen y = rnormal(50, 9)
postfile buffer meanhat1 meanhat2 using sampmeans2, replace
forvalues i=1/100 {
   preserve
   sample 30, count
   qui sum y
   scalar m1 = r(mean)
   scalar m2 = (y[1] + y[30])/2
   *di m1 m2
   post buffer (m1) (m2)
   restore
}
postclose buffer
```

Efficiency

```
use sampmeans2, clear
sum meanhat1 meanhat2
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    meanhat1 |        100    50.20508    1.498602    46.9187   53.75912
    meanhat2 |        100    49.38842    5.761486   33.89989   63.43159
```



Efficiency

- Both are unbiased; the expected value of the estimator is very close to 50. But the **standard error (SE)** of the estimators are different
- The SE of the usual estimator for the mean $(\sum \frac{x_i}{n})$ is smaller than the estimator $\frac{y[1]+y[30]}{2}$, the mean of the first and last observations
- Moral of the story, you prefer the estimator that has smaller variance. Just keep in mind that **both estimators should be unbiased**. No point in preferring a estimator with smaller variance if it's biased
- **Reminder**: Simulations a are not proofs, bu they help you get a sense of what you should find in theory

Consistency

- This property of an estimator is more intuitive and an important requirement of any estimator
- A good estimator should **get closer to the the true parameter the larger the sample** size
- This is not true for the example of y_n as an estimator of μ; the variance is always σ^2 regardless of the sample size
- Bias and efficiency are properties given a sample size. Now we are concerned about the **behavior of an estimator when we make the sample size larger and larger**
- That's an **asymptotic property**

Consistency

- From Wooldridge: Let W_n be an estimator of θ based on a sample $Y_1, Y_2, ..., Y_n$ of size n. Then, W_n is a consistent estimator of θ if for every $\epsilon > 0$, $P(|W_n - \theta| > \epsilon) \to 0$ as $n \to \infty$
- In words, the **probability that the estimator W_n is different** than θ **gets smaller** and smaller **as the sample gets larger** and larger (remember, the population is infinitely large)
- That ϵ in the definition means that the previous statement is true regardless of how large of small one sets the difference between W_n and θ to be
- It's not the the estimator is going to **always be closer** the larger the sample size gets (there is randomness) but **it is likely that it will**

Consistency

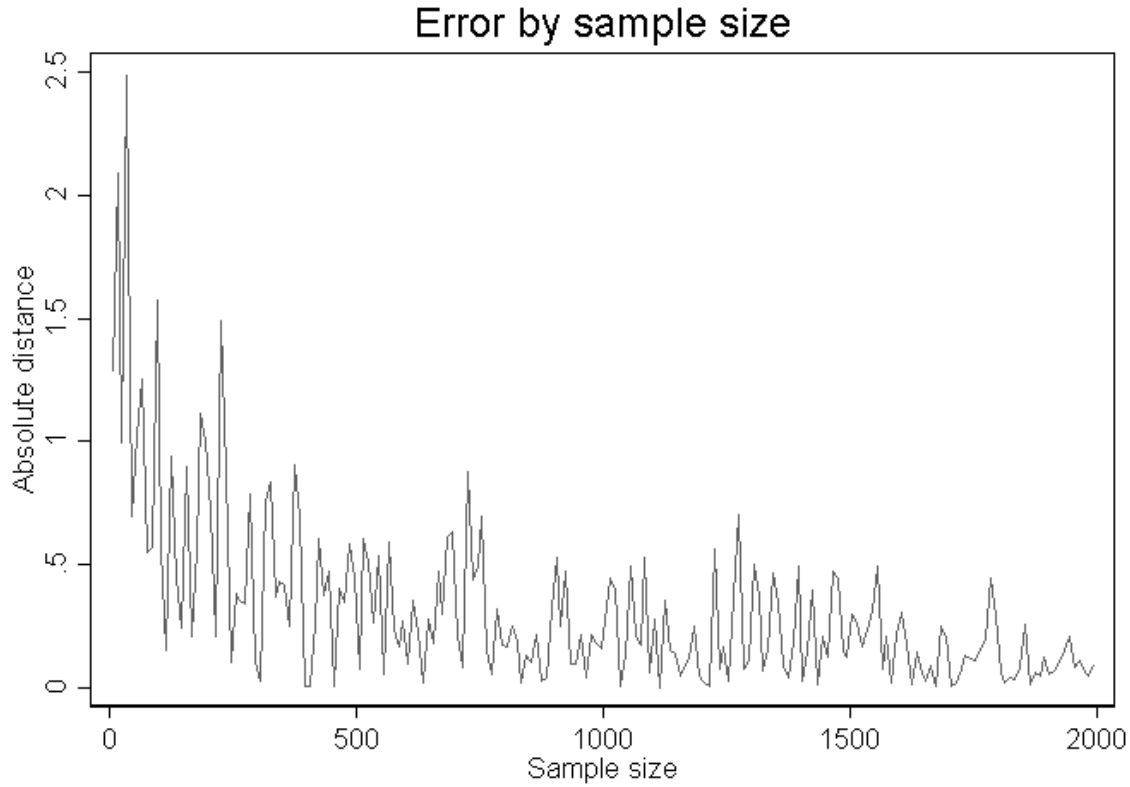
- Note that now we need to use use the "i" variable for the loops
- We are saving the mean and the sample size. The smaller sample size is 5, the larger 2000. We still have our population of 50,000 observations from $N(50, 9)$ (see do file)

```
postfile buffer meanhat1 samp using sampmeans3, replace
forvalues i= 5(10)2000 {
   preserve
   sample ‘i’, count
   qui sum y
   scalar m1 = r(mean)
   scalar s = r(N)
   post buffer (m1) (s)
   restore
}
postclose buffer
```

Consistency

- Note how in some cases the error was *larger* with a larger sample size. Remember, the **limit is in probability**

```
use sampmeans3, clear
gen dis = abs(meanhat1-50)
line dis samp, sort title("Error by sample size") ytitle("Absolute distance")
     xtitle("Sample size") saving(consi.gph, replace)
```



Loose ends

- Other probability distributions will show up when we do hypothesis tests: t-student, Chi-square, and the F-distribution
- Have this in mind:
- **Chi-square**: The Chi-square is the square of a standard normal (so only takes positive values, like the variance)
- **F distribution**: The F distribution originates from a ratio of two Chi-square distribution (like a ratio of variances)
- **t-student**: The t-student looks like a normal distribution with fatter tails and originates from a ratio of a normal to a Chi-square distribution

Summary

- Remember calculus and basic statistics. Learn about conditional densities, conditional expectation, rules of expected values, variances
- We have a model of how to think about a population, a sample, and an estimator
- A good estimator has three basic properties
- Make sure you understand and remember each property; it will help you understand Methods II
- Go over the simulations and make your own simulations