

Chapter 5 – Limit Theorems

def'n: A sequence of random variables are *independent and identically distributed (iid)* if they are independent and all have the same marginal distribution.

Example: $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \exp(\theta)$ means that the X_i are mutually independent and they all have an $\exp(\theta)$ distribution. The joint pdf of (X_1, \dots, X_n) is

$$f(x_1, \dots, x_n) = \prod_{i=1}^n \theta e^{-\theta x_i},$$

on $R_{\mathbf{X}} = (0, \infty) \times \dots \times (0, \infty)$.

Where are we going?

Here's a snapshot of why probability and random variables are useful ideas and where we are going.

Example: Diabetes is a growing concern among Inuit, who as a people have gradually shifted to a diet rich in carbohydrates from their traditional diet of seafood. A random sample of $n = 220$ Inuit from Labrador (part of a Canadian province) yielded 17 diabetics.

We would like to show that the prevalence of diabetes is higher among Inuit in Labrador than the rest of Canada, where the prevalence has been estimated previously to be 0.05. Do we have sufficient evidence to refute the status quo $H_0 : \pi = 0.05$?

We can recast this in terms of random variables. Let X_1, X_2, \dots, X_{220} be random variables such that $X_i = 1$ if subject i has diabetes and $X_i = 0$ if not. Let π be the unknown proportion of Inuit in Labrador with diabetes (we'd have to test every Inuit in Labrador to find π exactly). Then

$$X_1, \dots, X_{220} \stackrel{iid}{\sim} \text{Bern}(\pi).$$

We don't know π , but we do have $\sum_{i=1}^{220} x_i = 17$. Once the X_i are observed, we denote them x_i because they're no longer random. We can estimate π by $\hat{\pi}(x_1, \dots, x_{220}) = \hat{\pi}(\mathbf{x}) = \bar{x}_{220} = 17/220 \approx 0.077$. This is both the maximum likelihood estimate (MLE) and the method of moments (MOM) estimate. More on these later.

We certainly have evidence that $\pi > 0.05$, but how *strong* is this evidence? That is, how unlikely is seeing $\hat{\pi} = 0.077$ if $H_0 : \pi = 0.05$ is really true? More on this later too.

The main result in this chapter is that *sample means estimate population means* and *sample standard deviations estimate population standard deviations*.

Let $X_1, X_2, \dots, X_{n-1}, X_n, \dots$ be *iid* from any distribution. Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } \hat{\sigma}_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

These are the *sample mean* and *sample standard deviation*, obtained from the *sample variance*.

The Law of Large numbers tells us that when n is large \bar{X}_n is close to $\mu = E(X_i)$ with high probability, and $\hat{\sigma}_n$ is close to $\sigma = \sqrt{\text{var}(X_i)}$ with high probability.

Two properties of sample means

Let X_1, X_2, \dots, X_n be *iid* with $\mu = E(X_i)$ and $\text{Var}(X_i) = \sigma^2$. Then

$$E(\bar{X}_n) = \mu \text{ and } \text{Var}(\bar{X}_n) = \sigma^2/n.$$

1. Since $E(\bar{X}_n) = \mu$ we say \bar{X}_n is an *unbiased* estimator of μ . \bar{X}_n tends to on average “get it right.”
2. Since $\text{Var}(\bar{X}_n) = \sigma^2/n$, the variability of \bar{X}_n decreases as n gets large. Not only does \bar{X}_n typically get it right, but as we collect more data (n gets big), we’re *more likely* to get it right.

These are useful properties of \bar{X}_n . However, we can state things a bit more formally and strongly, via the LLN and CLT...

5.2 Law of large numbers

def'n: Let S_n be a sequence of random variables. We say S_n converges in probability to the number a if for every $\epsilon > 0$ we pick, $\lim_{n \rightarrow \infty} P(|S_n - a| > \epsilon) = 0$. We write this $S_n \xrightarrow{P} a$.

The idea is that we can pick a really small positive number ϵ , and the probability that S_n is at least ϵ away from a goes to zero. In other words, when n gets large, there's a high probability that S_n is near a .

Example: Let $S_n \sim \exp(n)$. Then $S_n \xrightarrow{P} 0$. We'll show this with a picture and also prove it formally using the cdf $F_n(x) = 1 - e^{-nx}$.

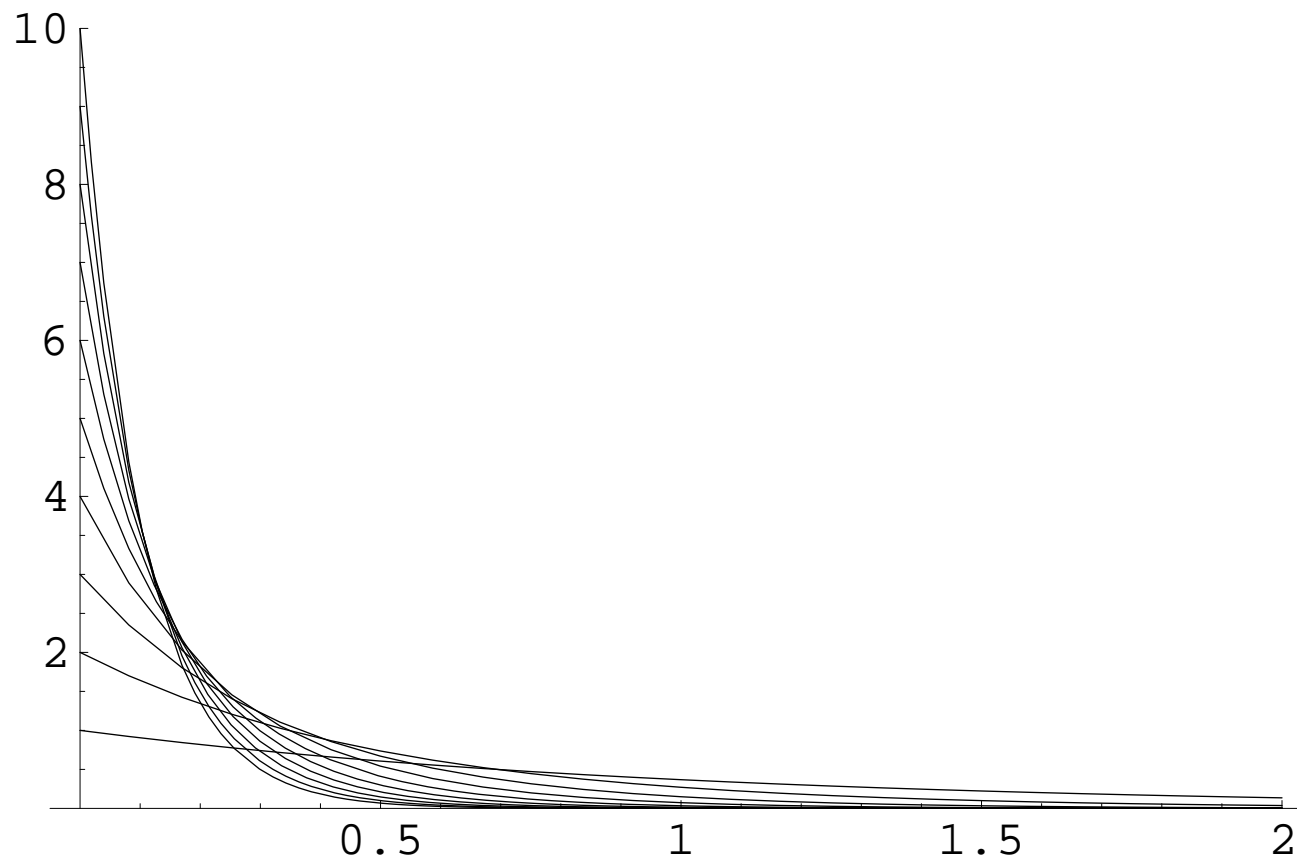


Figure 1: $\text{exp}(n)$ densities for $n = 1, \dots, 10$.

Example: Let $S_n \sim N(5, 1/n^2)$. Then $S_n \xrightarrow{P} 5$.

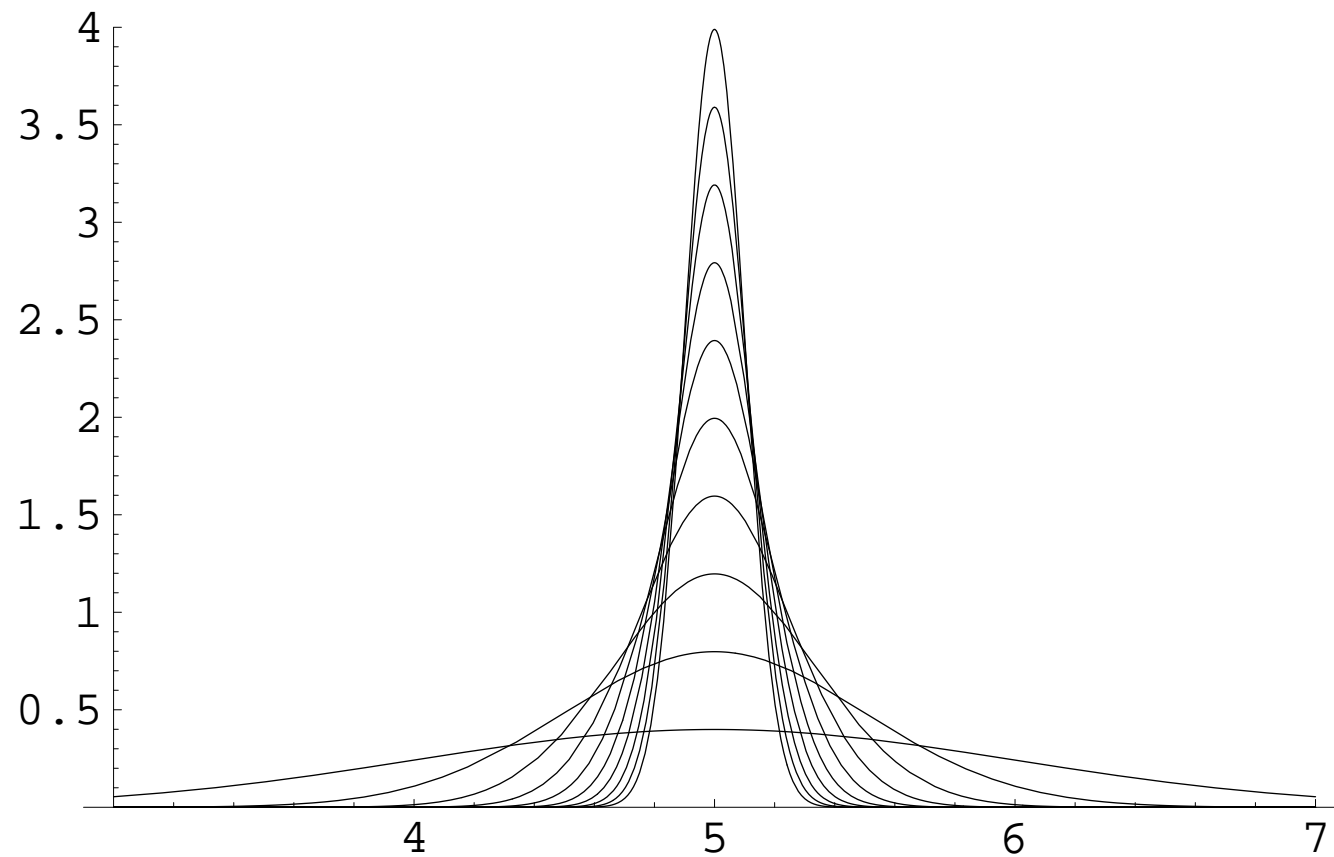


Figure 2: $N(5, 1/n^2)$ densities for $n = 1, \dots, 10$.

Law of large numbers: Let X_1, X_2, X_3, \dots be *iid* from **any** distribution with $\mu = E(X_i)$ and $\sigma^2 = \text{Var}(X_i)$. Let \bar{X}_n be the random sample mean $\bar{X}_n = (X_1 + \dots + X_n)/n$. Then $\bar{X}_n \xrightarrow{P} \mu$. We say “ \bar{X}_n is consistent for μ .” Consistency is an important criterion for any estimator.

We are not assuming anything about the distribution of the X_i 's! Yet we can estimate the unknown μ with the sample mean \bar{X}_n . As n gets large, the probability that \bar{X}_n is close to μ goes to one.

Example: Let's simulate some data. On the next slide one set of observed $\bar{X}_n = \bar{x}_n$ are plotted for $n = 1000, 2000, 3000, \dots, 100000$ for *iid* $\text{exp}(0.1)$ data. Note here that $\mu = E(X_i) = 10$.

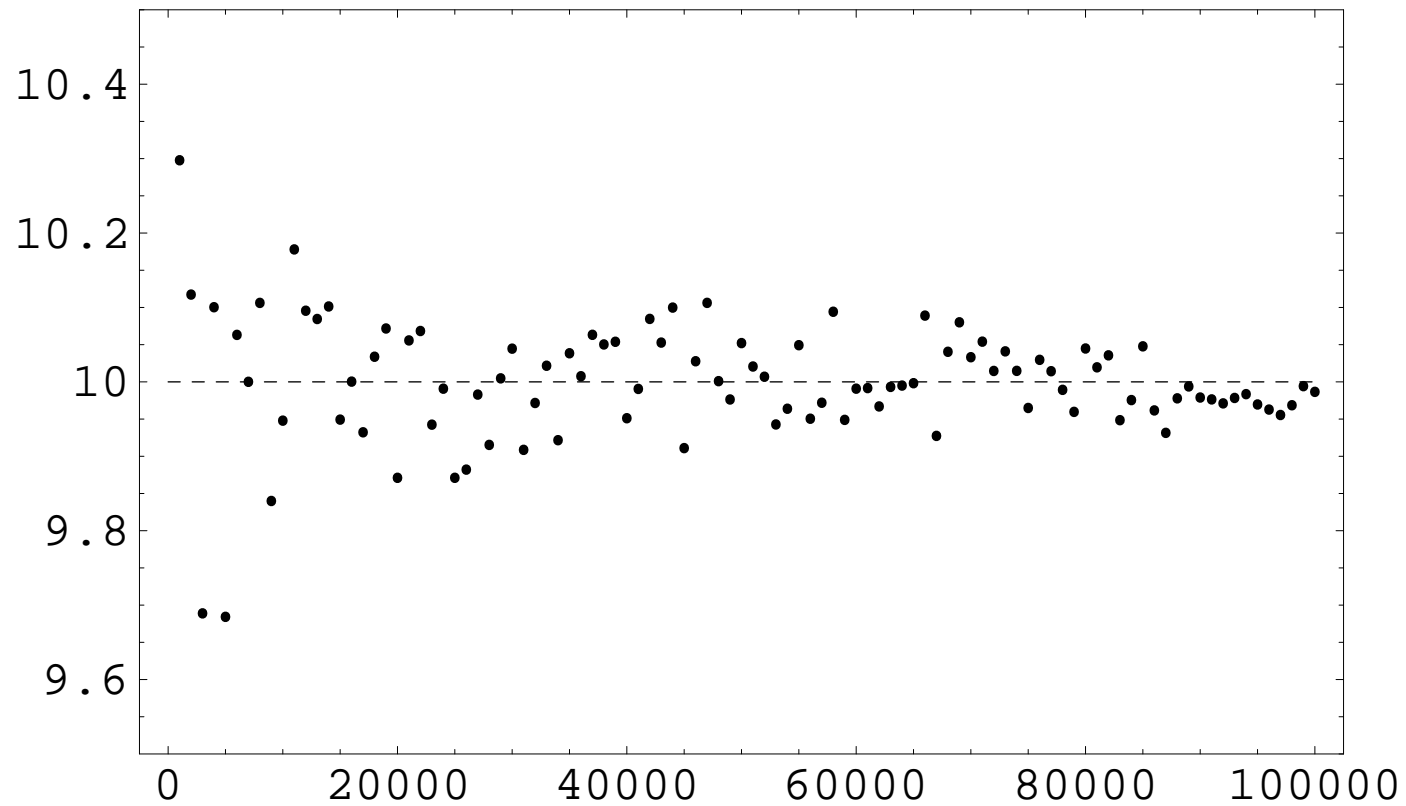


Figure 3: \bar{x}_n from $X_1, \dots, X_n \stackrel{iid}{\sim} \exp(0.1)$ versus n .

Using a bit of math, one can also show that **sample variances** and **sample standard deviations** consistently estimate their population counterparts:

$$\hat{\sigma}_n \xrightarrow{P} \sigma \text{ and } \hat{\sigma}_n^2 \xrightarrow{P} \sigma^2 = \text{var}(X_i).$$

As discussed in the context of boxplots, **sample percentiles** also consistently estimate their population counterparts:

$$\hat{x}_p = x_{(\lceil np \rceil)} \xrightarrow{P} x_p,$$

where x_p satisfies $P(X_i \leq x_p) = p$. This implies, for example, that the sample interquartile range, denoted $\widehat{\text{IQR}} = x_{(\lceil n0.75 \rceil)} - x_{(\lceil n0.25 \rceil)}$ consistently estimates the true, unknown interquartile range $\text{IQR} = F^{-1}(0.75) - F^{-1}(0.25)$ for continuous distributions.

Let's compare sample estimates of the IQR for $X_1, X_2, \dots \stackrel{iid}{\sim} N(0, 1)$ to the truth.

```
> n=10; d=rnorm(n,0,1); s=sort(d); s[ceiling(n*0.75)]-s[ceiling(n*0.25)]
[1] 0.8946914
> n=100; d=rnorm(n,0,1); s=sort(d); s[ceiling(n*0.75)]-s[ceiling(n*0.25)]
[1] 1.527814
> n=1000; d=rnorm(n,0,1); s=sort(d); s[ceiling(n*0.75)]-s[ceiling(n*0.25)]
[1] 1.351120
> n=10000; d=rnorm(n,0,1); s=sort(d); s[ceiling(n*0.75)]-s[ceiling(n*0.25)]
[1] 1.345663
> n=100000; d=rnorm(n,0,1); s=sort(d); s[ceiling(n*0.75)]-s[ceiling(n*0.25)]
[1] 1.344759
> n=1000000; d=rnorm(n,0,1); s=sort(d); s[ceiling(n*0.75)]-s[ceiling(n*0.25)]
[1] 1.349091
> qnorm(0.75,0,1)-qnorm(0.25,0,1) # HERE'S THE TRUE IQR!
[1] 1.348980
```

It takes awhile, but consistency eventually kicks in and we approximate the true (but usually unknown!) IQR quite well to 4 decimal places.

Let's try this out using sample means and standard deviations:

```
> n=10; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= 0.2049119 sample std= 0.921117
> n=100; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= -0.06257447 sample std= 1.077131
> n=1000; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= -0.02328674 sample std= 1.055744
> n=10000; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= 0.01103710 sample std= 0.9961411
> n=100000; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= 0.000171471 sample std= 1.001763
> n=1000000; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= -0.001451602 sample std= 1.000221
> n=10000000; d=rnorm(n,0,1); cat("sample mean=",mean(d),"sample std=",sqrt(var(d)),"\n")
sample mean= 0.0001992165 sample std= 1.000292
```

We know that exactly, $\mu = E(X_i) = 0$ and $\sigma = \sqrt{\text{var}(X_i)} = 1$. It takes large sample sizes to get these down to 4 decimal places.

What does this tell you about papers that report sample estimates to 4, 5, or 6 decimal places when $n = 57$?

5.3 Convergence in distribution and the Central Limit Theorem

We can say even more about \bar{X}_n as an estimator for $\mu = E(X_i)$.

The **Law of Large Numbers** tells us \bar{X}_n is close to μ as n gets big.

The **Central Limit Theorem** tells us that \bar{X}_n is approximately normal *regardless of the distribution on the X_1, \dots, X_n .*

def'n: Let S_n be a sequence of random variables. We say S_n converges in distribution to S if $\lim_{n \rightarrow \infty} P(S_n \in A) = P(S \in A)$ for all sets A . We write this $S_n \xrightarrow{D} S$.

If $S_n \xrightarrow{D} S$ then for large n we can approximate anything having to do with S_n (probabilities, $E(S_n)$, $\text{Var}(S_n)$, quantiles, etc.) using S instead.

Central Limit Theorem (CLT): Let X_1, X_2, X_3, \dots be *iid* from any distribution with mean $\mu = E(X_i)$ and variance $\text{Var}(X_i) = \sigma^2$. Then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$.

Here, $\sqrt{n}(\bar{X}_n - \mu)$ is a function of random \bar{X}_n and fixed μ .

Recall that for any $Y \sim N(\mu, \sigma^2)$, $bY \sim N(b\mu, b^2\sigma^2)$. So $\sqrt{n}(\bar{X}_n - \mu) \overset{\bullet}{\sim} N(0, \sigma^2)$ implies that $(\bar{X}_n - \mu) \overset{\bullet}{\sim} N(0, \sigma^2/n)$. Also recall that $a + Y \sim N(\mu + a, \sigma^2)$ so $\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$.

CLT restated: Let X_1, X_2, X_3, \dots be *iid* from any distribution with mean $\mu = E(X_i)$ and variance $\text{Var}(X_i) = \sigma^2$. Then $\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$ where “ $\overset{\bullet}{\sim}$ ” is read “approximately distributed as.”

Wait! The CLT states that $\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$. But we already know, *exactly*, that

$$E(\bar{X}_n) = \mu \quad \text{and} \quad \text{var}(\bar{X}_n) = \frac{\sigma^2}{n},$$

you say. And this is true *for any size* n .

That's right. Alls the CLT does is add the “normality” part! But that's a big deal, because once we have a *sampling distribution* for \bar{X}_n in terms of the unknown μ and σ , then we can do interesting things like (a) find a plausible interval that the unknown μ lies in based on \bar{X}_n and $\hat{\sigma}_n$, and (b) test whether μ is equal to some hypothesized value, like $\mu = 17.5$ calories/hour.

- The CLT is proved in your text using *moment generating functions* and Taylor's Theorem from calculus (p. 184). It's not hard to prove, but not very illuminating either. If we have time, we'll carefully prove it later.
- The CLT is perhaps the most important result in statistics. It does not care if the X_1, X_2, X_3, \dots are continuous, discrete, or neither. It generalizes to random vectors. It generalizes to X_1, X_2, X_3, \dots that are independent but not necessarily identically distributed. It is used to prove that MOM and MLE estimators are normal.
- The normal approximation gets better with large n , but how big does n need to be? It depends on the true, typically unknown distribution of the X_1, X_2, X_3, \dots . The more “non-normal looking” the pdf or pmf is, the larger n needs to be.

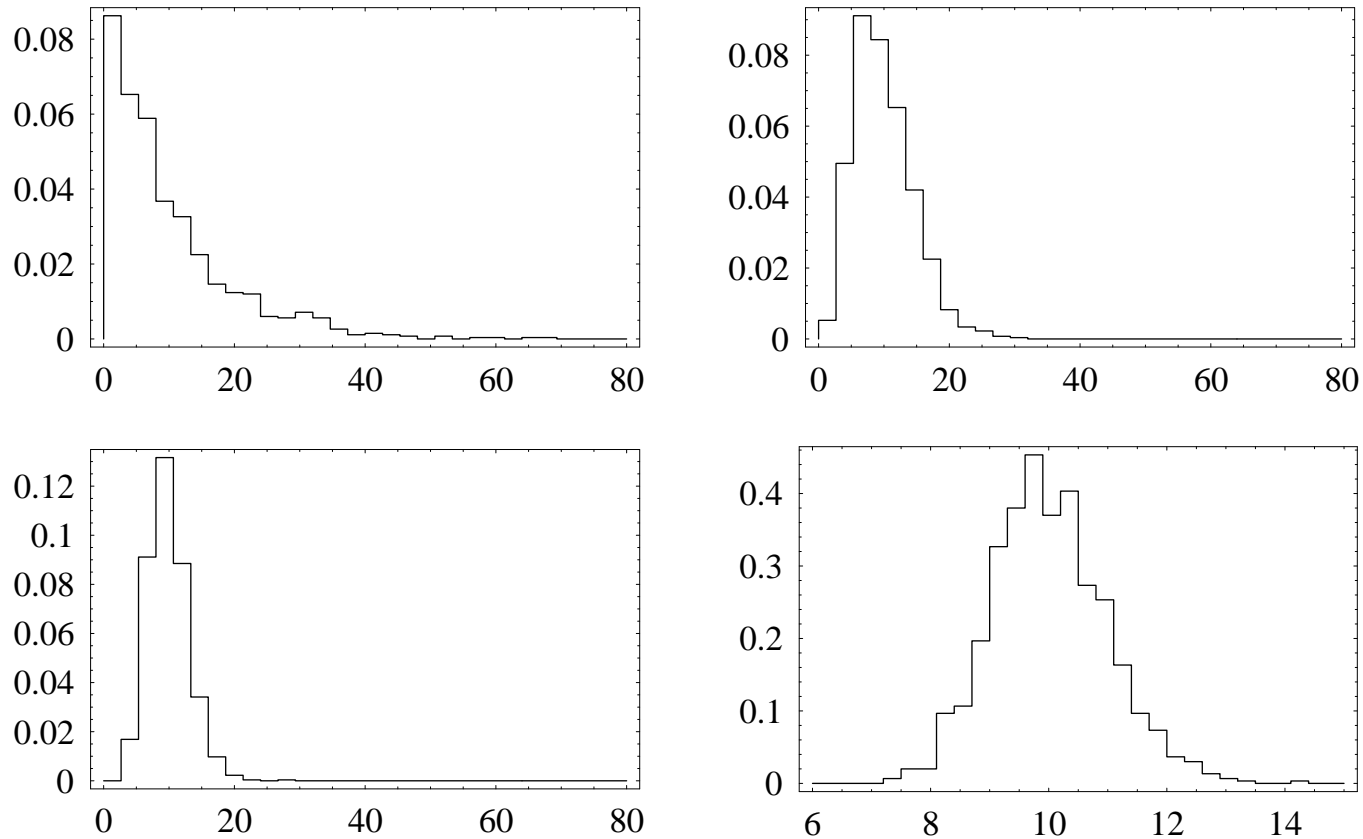


Figure 4: Histogram of 1000 \bar{X}_n 's from simulated $\text{exp}(0.1)$ data, $n = 1, 5, 10, 100$. The CLT really “kicks in” around $n = 10$.

Better living through simulation

If we are interested in aspects of a density (or pmf) $f(x)$ and can *simulate* values from it, we can estimate summaries such as the mean, variance, and quantiles by simulating a lot of *iid* data and computing the sample estimates.

For example, say $f(x) = \pi \cos(\pi x/2)/4$ on $-1 \leq x \leq 1$, where here $\pi = 3.14159\dots$

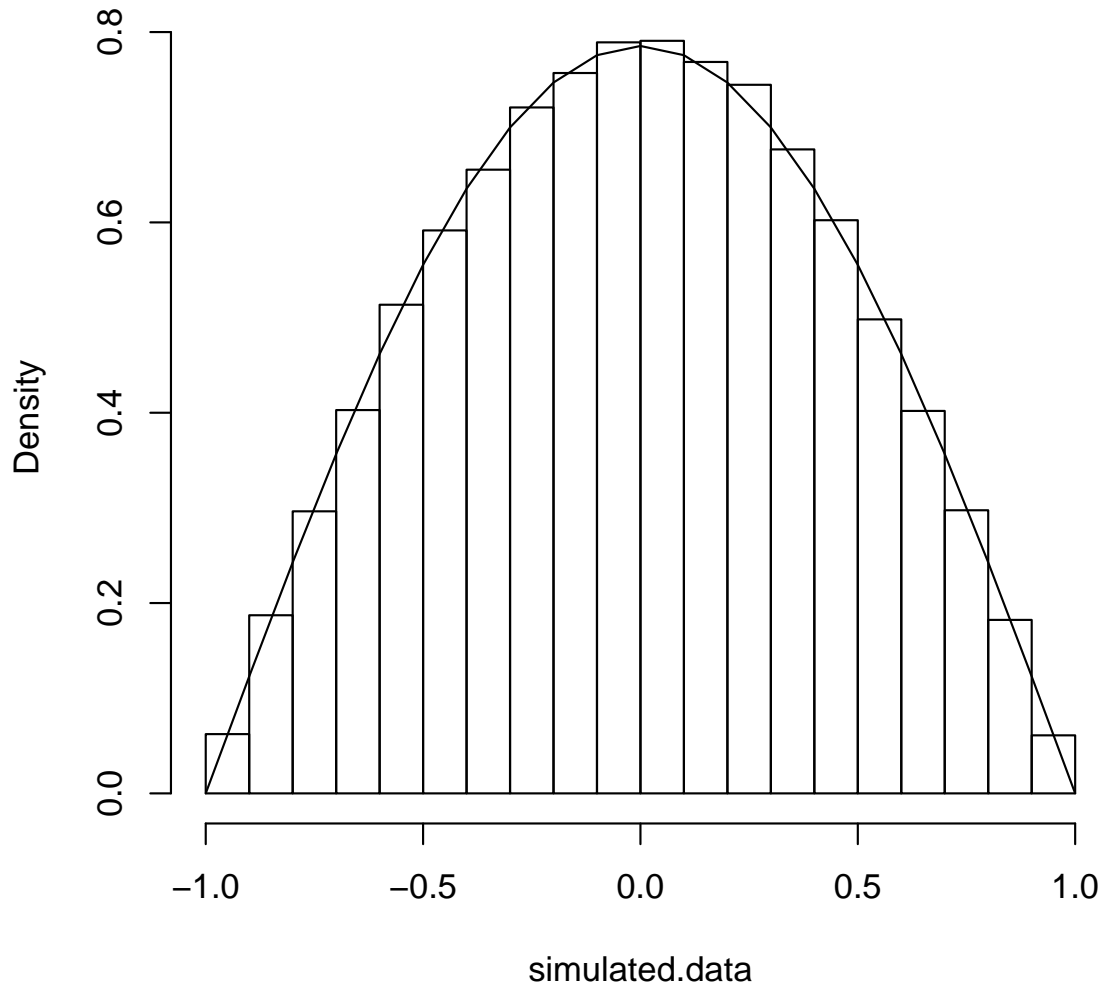
Through calculus we can show $F(x) = [1 + \sin(\pi x/2)]/2$ and $F^{-1}(p) = -2 \arcsin(1 - 2p)/\pi$ where $\arcsin(x) = \sin^{-1}(x)$.

Then, *exactly*, $x_{0.25} = -0.33333\dots$ and $x_{0.5} = 0$ and $x_{0.75} = 0.33333\dots$
Also, $E(X) = 0$ and $\text{sd}(X) = \sqrt{\pi^2 - 8}/\pi = 0.435236\dots$

We can instead simulate all of these summaries using the inverse cdf method. If $U \sim U(0, 1)$, then $X = F^{-1}(U)$ is distributed with density $f(x)$.

```
> f=function(x){pi*cos(x*pi/2)/4}
> x=seq(-1,1,0.1); y=f(x)
> Finv=function(p){-2*asin(1-2*p)/pi}
> simulated.data=Finv(runif(100000))
> hist(simulated.data,freq=F)
> lines(x,y)
> mean(simulated.data)
[1] 0.000629261
> sd(simulated.data)
[1] 0.4340932
> quantile(d,0.25)
      25%
-0.3351720
> quantile(d,0.75)
      75%
0.3342226
```

Histogram of simulated.data



You will have a homework problem where you verify that the simulated and true values coincide. The density is $f(x) = 3x^2$ on $0 \leq x \leq 1$.

Method of composition

A hierarchical model lists random variables in stages.

Recall the apple/rainfall example. Say X is the amount of rain in inches an apple tree gets over a summer. Assume $X \sim \exp(0.1)$.

Conditional on $X = x$, the yield of apples in pounds is $Y|X = x \sim \exp\left(\frac{1}{2x}\right)$. This is written hierarchically as

$$Y|X \sim \exp(0.5/X), \quad X \sim \exp(0.1).$$

This implicitly defines a joint distribution on (X, Y) .

Say we are really only interested in Y . We can simulate from the marginal distribution of Y by simply taking

$$X \sim \exp(0.1), \quad \text{then } Y \sim \exp(0.5/X).$$

```
> x=rexp(100000,0.1) # simulate from X
> y=rexp(100000,0.5/x) # simulate from Y|X
> quantile(y,0.5)      # estimate median of Y, etc...
      50%
7.831159
> quantile(y,0.75)-quantile(y,0.25)
      75%
20.58463
> mean(y)
[1] 19.96257
> sd(y)
[1] 34.63597
```

One final, important use for simulation. Probabilities are expectations,

$$P(X \in A) = E [I\{X \in A\}].$$

So a probability can be estimated by simply counting the number of X_1, \dots, X_{100000} in A . Say we want to estimate the probability of getting more than 10 lbs of apples from a tree. `y[(y>10)]` gives a new vector with only those elements in `y` that are larger than 10. Then we just need to count how many elements are in this new vector and divide by the number of elements in `y`.

```
> length(y[(y>10)])/length(y)
[1] 0.44252
```

Let's try and make this more transparent with a small sample size of 10:

```
> x=rexp(10,0.1)
> y=rexp(10,0.5/x)
> x
 [1]  1.006076  4.440620 18.154714 12.547808  5.640018 12.564547  4.830029
 [8]  1.258576  9.335616  2.341031
> y
 [1]  0.7108791  5.9083802 16.1061430  4.4554008  7.2603352 13.9726632
 [7] 23.8602768  1.3610906 17.4728531  0.4386384
> mean(y)
 [1] 9.154666
> sd(y)
 [1] 8.173125
```

Wow! That's much different from when we used 100,000 as a simulation size.

```
> y[(y>10)]
[1] 16.10614 13.97266 23.86028 17.47285
> length(y[(y>10)])
[1] 4
> length(y)
[1] 10
> greater=y[(y>10)]
> greater
[1] 16.10614 13.97266 23.86028 17.47285
> length(greater)
[1] 4
```

What happens when we repeat with 10, then with 100,000?

```
> x=rexp(10,0.1)
> y=rexp(10,0.5/x)
> mean(y)
[1] 27.43546
> x=rexp(100000,0.1)
> y=rexp(100000,0.5/x)
> mean(y)
[1] 20.07581
```