ntrial = 64
xbar = numeric(ntrial)
par(mfrow=c(8,8))
for( trial in 1:ntrial ){
    x = rnorm(50, 0, 1)
    hist(x, breaks=10)
    xbar[trial] = mean(x)
}
hist(xbar, main=""")

Q: What's $\bar{x}$ in each hist above?
Q: What's the mean of the $\bar{x}$'s?
Q: What's $s$ in each hist above?
Q: What's $s$ of the $\bar{x}$'s?
Lecture 15 (Ch. 5)

Sampling Distribution: Extremely Important!!!

Population, \( x \)

\( \mu_x, \sigma_x, \text{median} \)

Sample 1

\( \bar{x}, s, p \)

Sample 2

\( \bar{x}, s, p \)

Sample 10

\( \bar{x}, s, p \)

Sample prop.

\( \Rightarrow 10^6 \bar{x}'s \Rightarrow \text{histogram} \)

\( \sim N(\mu, \sigma^2) \)

2 important quantities:

- One estimates the pop. parameter (e.g., \( \mu \)), the other tells us how certain that estimate is.

Precise

Statistic (\( \bar{x} \))

\( \sim N(\mu, \sigma^2) \)

or \( S \)

or \( p \)

Random vars.

The sampling dist. (of the sample mean) is a distribution, i.e., a p(x) or an f(x) that can be derived mathematically, or simply assumed as a description of the population of all \( \bar{x}'s \).

The only reason I talk about a histogram is to make the concept of the sampling dist. more intuitive. The histogram is sometimes called the "empirical sampling dist."
Note that the sampling distribution is the distribution of a sample statistic. For example, the sample distribution of the sample mean, tells us how the sample means are distributed. Similarly, the sample distribution of the sample proportion, tells us how the sample proportions are distributed. Etc.

What is the sampling distribution of $\bar{X}$? Normal, Poisson, ...?

Later!

But even without knowing the distribution, we can still find its mean ($E[\bar{X}]$ or $\mu_{\bar{X}}$) and variance ($V[\bar{X}]$ or $\sigma^2_{\bar{X}}$).

If the population (i.e., distribution) has mean $\mu_x$ and std. dev. $\sigma_x$, then the mean of the sampling distribution of sample mean ($\mu_{\bar{X}}$): $\text{Std. dev.}$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ $\ldots$ ($\sigma_{\bar{X}}$):

$$\mu_{\bar{X}} = E[\bar{X}] = \mu_x$$

proof, below.

$$\sigma_{\bar{X}} = \sqrt{V[\bar{X}]} = \frac{\sigma_x}{\sqrt{n}}$$

sample size

Sometimes called "standard error of mean."
Derivation: Suppose we do not know the distr. of the population (p(x), f(x)), but we do know its \( \mu_x \) and \( \sigma_x \).

Of course, if you do know the pop. distr., then you can compute \( \mu_x, \sigma_x \) as before:

\[
E(x) \equiv \mu_x = \frac{\Sigma x \cdot p(x)}{x} \quad (\text{or } \int x \cdot f(x) \, dx) \\
V(x) = \sigma_x^2 = \Sigma \left( x - \mu_x \right)^2 p(x) \quad (\text{or } \int (x - \mu_x)^2 \, dx)
\]

Recall, \( E(ax) = a \cdot E(x) \), \( V(ax) = a^2 \cdot V(x) \), \( a = \text{constant} \). Then

\[
\mu_x = E(\bar{X}) = E\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \left( \sum_i E[X_i] \right) = \frac{1}{n} \cdot \mu_x \cdot \frac{1}{n} = \mu_x. \\
\]

The \( i^{\text{th}} \) obs. is a random value. There is nothing special about the \( i^{\text{th}} \) obs.

So, just drop the "\( i \)". Then \( E[X_i] = E[X] = \Sigma x \cdot p(x) = \mu_x \).

Alternatively, work out \( E[x_i] \) for each \( i \), e.g., \( i = 1 \)

\[
E[X_i] = \frac{\Sigma x_i \cdot p(x_i)}{x_i} = \mu_x, \quad E[X_i] = \mu_x, \quad \forall i.
\]

\[
\sigma_x^2 = V(\bar{X}) = V\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \left( \frac{1}{n} \right)^2 \sum \left( \frac{1}{n} \right) \cdot V[X_i] = \left( \frac{1}{n} \right)^2 \cdot \sigma_x^2 \Rightarrow \sigma_x^2 = \frac{\sigma_x^2}{n} \quad \Rightarrow \quad \sigma_x = \sqrt{V(x)} = \frac{\sigma_x}{\sqrt{n}}
\]

The Var. of each element in the pop.

The Var. of the pop.
In summary:

\[ \mu_x = \mathbb{E}[x] = \mu \]

Tells us that we can use the sample mean
(from the one sample of size \(n\)) to estimate
The pop. mean \( \mu \) with accuracy. *(see box, below)*

\[ \sigma_x = \sqrt{\mathbb{V}[x]} = \frac{\sigma_x}{\sqrt{n}} \]

Tells us that the typical deviation in \( \bar{x} \) is \( \frac{\sigma_x}{\sqrt{n}} \),
and so it tells us how precise
is our estimate of \( \mu_x \). *(see box, below)*

Note that \( \mu_x, \sigma_x, \mu_{\bar{x}}, \sigma_{\bar{x}} \) are means and std. dev.
of distributions, not of data. We are dealing
with distributions, even though the thought exp. involved a hist.

\[
\mu_x = \sum_x x \cdot f(x), \quad \int x f(x) \, dx \quad \sigma_x^2 = \sum_x (x - \mu_x)^2 f(x), \quad \int (x - \mu_x)^2 f(x) \, dx
\]

\( \bar{x} \) and \( s_x \) are measures of **Accuracy & Precision**:

\[ \mu_{\bar{x}} \quad \text{and} \quad \sigma_{\bar{x}} \]

\[ \text{Accuracy} \ (\bar{x} - \mu_x) \]

\[ \text{Precision} \]

\[ \text{std dev} \]

\[ \text{True/pop mean} \]

![Diagram](image)

Q: OK, so now we know \( \mathbb{E}[\bar{x}] \) and \( \mathbb{V}[\bar{x}] \), but what is
The distribution of \( \bar{x} \) itself?
We arrive at the Central Limit Theorem (CLT):

**Weak version:** If \( x \sim N(\mu, \sigma) \), then \( \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \)

**Strong version:** If \( x \sim \text{any dist. with mean } \mu \text{, var. } \sigma^2 \)

Then \( \bar{x} \sim N(\mu_x, \frac{\sigma^2}{n}) = \bar{x} \) for large \( n \).

In English: For any pop. with mean \( \mu \text{ and variance } \sigma^2 \), the sampling dist. of the sample means is normal with \( \mu = \mu \text{x, } \sigma = \frac{\sigma}{\sqrt{n}} \).

So, if we know the pop. (i.e. \( f(x), p(x) \)), then we can compute the prob. that a random sample mean will be somewhere.

E.g. \( \Pr(a < \bar{x} < b) \). This is how:

1) Compute \( \mu_x, \sigma \):
   \[ \mu_x = \mathbb{E}[x] = \sum_x x \cdot f(x), \quad \sigma^2 = \mathbb{V}[x] = \sum_x (x - \mu_x)^2 \cdot f(x). \]

2) From CLT we know
   \[ \bar{x} \sim N(\mu_x, \frac{\sigma^2}{n}). \]

3) Then standardize:
   \[ Z = \frac{\bar{x} - \mu_x}{\sigma / \sqrt{n}} \sim N(0, 1) \]

4) Finally
   \[ \Pr(a < \bar{x} < b) = \Pr\left( \frac{a - \mu_x}{\sigma / \sqrt{n}} < \frac{\bar{x} - \mu_x}{\sigma / \sqrt{n}} < \frac{b - \mu_x}{\sigma / \sqrt{n}} \right) \]

   \[ = \Pr\left( \frac{a - \mu_x}{\sigma / \sqrt{n}} < Z < \frac{b - \mu_x}{\sigma / \sqrt{n}} \right) \Rightarrow \text{Table 1} \]

Compare with what we did in Ch 12:

\[ \Pr(a < \bar{x} < b) = \Pr\left( \frac{a - \mu_x}{\sigma / \sqrt{n}} < Z < \frac{b - \mu_x}{\sigma / \sqrt{n}} \right) \]
Suppose a sample of size 25 yields \( \bar{x}_{\text{obs}} = 3, s = 1 \).
If the population is \( N(\mu = 2, \sigma = 1) \), what's the probability of getting an even larger sample mean? \( M_x = 2 \)

\[
\text{prob}(\bar{x} > \bar{x}_{\text{obs}}) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}
\]

\[
\text{prob}(\bar{x} > 3) = \text{prob}(z > \frac{3 - 2}{1/\sqrt{25}}) = \text{prob}(z > 5) \approx 0
\]

This small probability suggests that \( \mu = 2 \) is a bad assumption. In fact, we may even guess that \( \mu \) is greater than 2 (closer to 3).
We will formalize these qualitative conclusions below.

Recall "prob" = proportion of samples (of size \( n \)) taken from the population, in the long-run (e.g. out of \( 10^8 \) samples).

Prob works on random variables:

I.e. \( \text{prob}(a < \bar{x} < b) \) is computable. \( \checkmark \)
\[
\text{prob}(a < \bar{x}_{\text{obs}} < b) \text{ is not } \checkmark
\]

Note: in these calculations we are assuming we know the pop. But we don't. Intuitively, these probs give us a sense of how likely it would be to get a random sample mean somewhere, IF the pop is given. In ch. 7, 8 we will come up with 2 ways of turning things around to say something about pop. from data.
A sampling distribution (e.g. of the sample maximum, for samples of size 50 taken from a standard normal) use 5000 trials.

b) Then, repeat but for sample minimum.

Turn-in: The code, and the resulting 2 histograms.

Fri: These “distributions” arise naturally when one tries to model extreme events, e.g. The biggest storms, The strongest earthquakes, The brightest stars, The smallest forms of life, etc.

---

1. Write R code to produce the sampling distribution (The sample maximum, for samples of size 50 taken from a standard normal). Use 5000 trials.

2. Then, repeat but for sample minimum.

3. Write R code to take 5000 samples of size n=100 from an exponential distribution with parameter \( \lambda = 2 \), and plot a qq.plot of the 5000 means. Recall that if the qq.plot is a straight line, then the histogram of the sample means is normal. This will show that the sample dist. of sample means is normal, even when the pop. is not.

---

A sampling distribution (e.g. of the sample mean) is a distribution, not a histogram of observed sample means; the histogram of sample means discussed in class is just an intuitive way of thinking about the sampling distribution; technically, it’s called the *empirical* sampling distribution. Of course, if the number of trials is infinite, then the empirical sampling distribution (i.e., the histogram) approaches the distribution. Anyway, to show that the sampling distribution is truly a distribution (not a histogram), let’s derive one mathematically – no data at all.

Consider a population described by a Bernoulli random variable, i.e., \( x = 0,1 \), following the Bernoulli distribution, i.e., \( p(x) = p^x (1-p)^{1-x} \). Suppose we take samples of size 2.

a) Write down all the possible samples. Hint: there are only 4.

b) For each of the possible samples, compute the sample mean.

c) For each of the possible samples, compute the probability. Hint: Use Bernoulli.

d) Based on your answers to parts a-c, find the probability of each of the possible sample means.

Note: your answer to part d *is* the sampling distribution of the sample mean! Note that it’s not a histogram, but a real distribution.
A sample of size 36 from a Normal pop. yields $\bar{x} = 3.5, s = 1$.

**a)** Under the assumption that $\mu_x = 2.5, \sigma_x = 2$, what's the prob of a sample mean larger than the one observed?

**b)** Under the assumption that $\mu_x = 2.5, \sigma_x = 2$, what's the prob of a sample mean smaller than the one observed?

**c)** Under the assumption that $\mu_x = 3.5, \sigma_x = 2$, what's the prob of a sample mean larger than the one observed?

**d)** Under the assumption that $\mu_x = 3.5, \sigma_x = 2$, what's the prob of a sample mean smaller than the one observed?