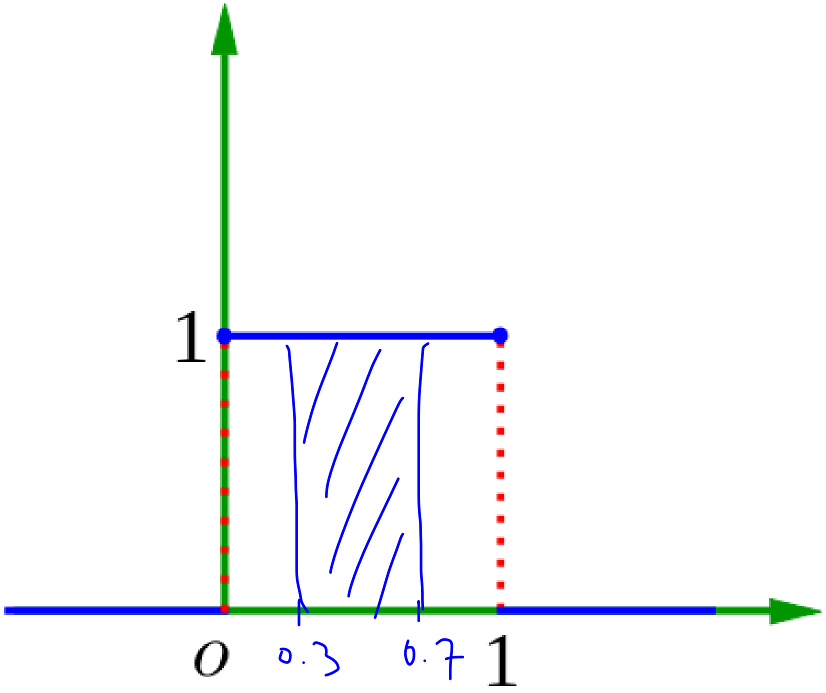
Lecture 8

Continuous Random Variables

Example: The random number generator will spread its output uniformly across the entire interval from 0 to 1 as we allow it to generate a long sequence of numbers. The results of many trials are represented by the density curve of a **uniform distribution**.



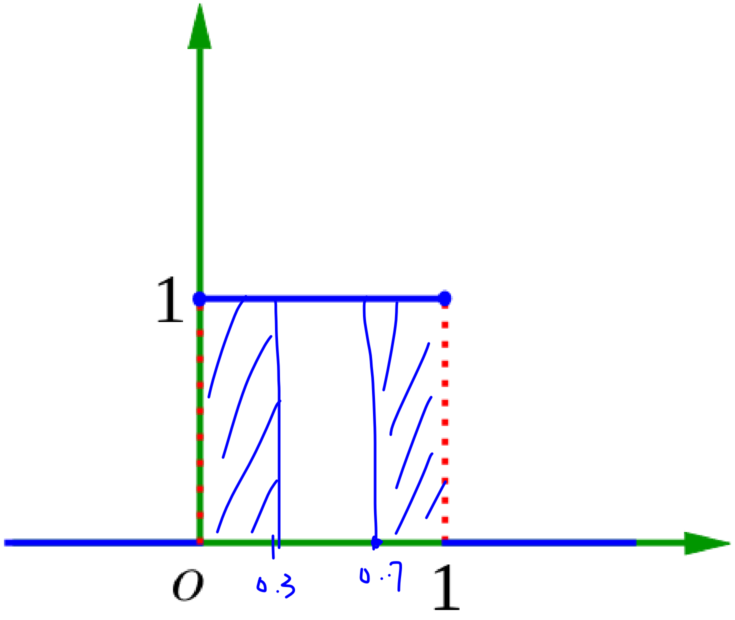
$1 \cdot (0.7 - 0.3)$

$X \in [0, 1]$

$P(0.3 \leq X \leq 0.7)$

$= 1 \cdot (0.7 - 0.3)$

$= 0.4$



$$P(X \le 0.3 \text{ or } X \ge 0.7)$$

$$= P(X \le 0.3) + P(X \ge 0.7)$$

$$= (0.3 - 0) + (1 - 0.7)$$

$$= 0.6$$

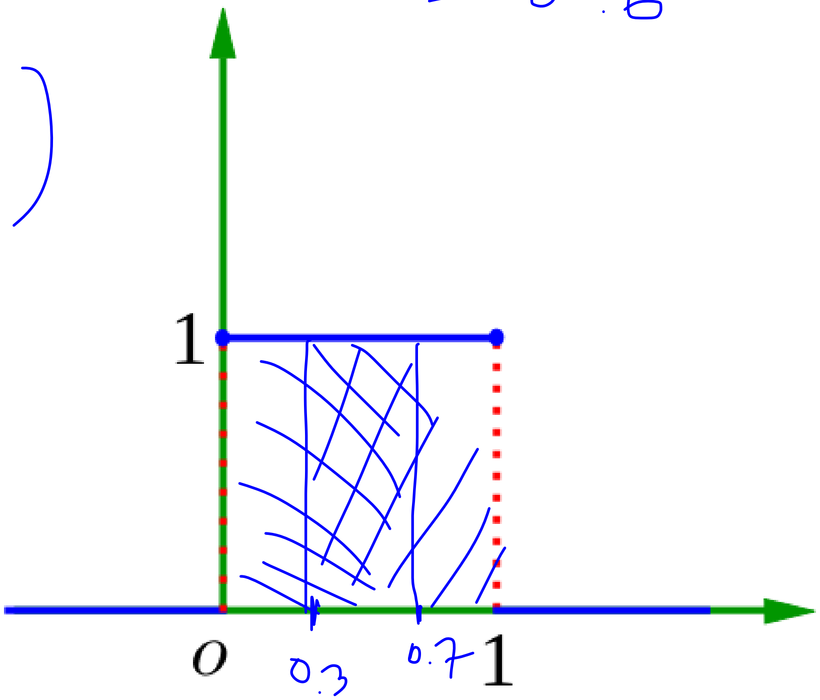
$$P(X \ge 0.3 \text{ or } X \le 0.7)$$

$$= P(X \ge 0.3) + P(X \le 0.7)$$

$$- P(0.3 \le X \le 0.7)$$

$$= (1 - 0.3) + (0.7 - 0) - (0.7 - 0.3)$$

$$= 0.7 + 0.7 - 0.4 = 1$$



- A **continuous random variable** *X* takes all values in an interval of numbers.

- The **probability distribution** of *X* is described by a density curve.

- The probability of any event is the **area** under the density curve and above the values of *X* that make up the event.

Note: $P(X = 0.7) = 0$

$$P(X \geq 0.7) = P(X > 0.7)$$

why?

$$P(0.69 \leq X \leq 0.71) = 0.02$$

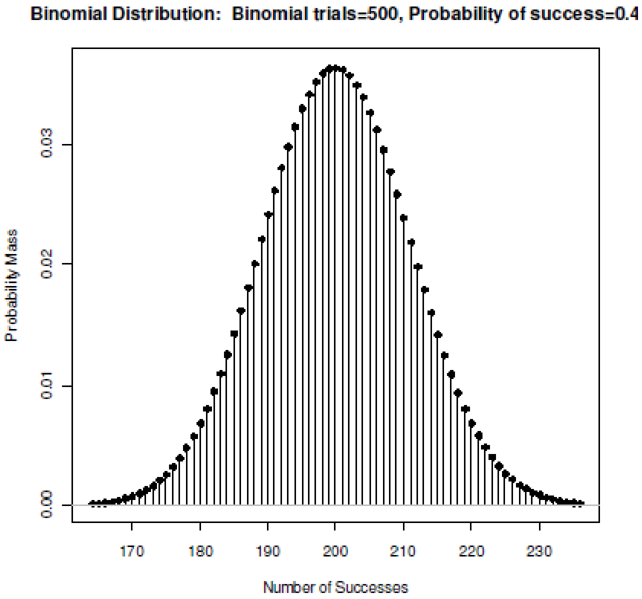
$$P(0.699 \leq X \leq 0.701) = 0.002$$

$$P(\underset{\substack{\downarrow \\ P(X=0.7)}}{0.699999 \leq X \leq 0.700001}) = \underset{\substack{\downarrow \\ \text{zero}}}{0.000002}$$

Normal Distributions

Let's look at the examples from the previous lecture:

Example: binomial distribution, n = 500, p = 0.4



X = counts

$X \sim Bin(n, p)$

$\mu_X = np$

$\sigma_X = \sqrt{np(1-p)}$

$= \sqrt{npq}$

$\hat{p} = \frac{X}{n}$

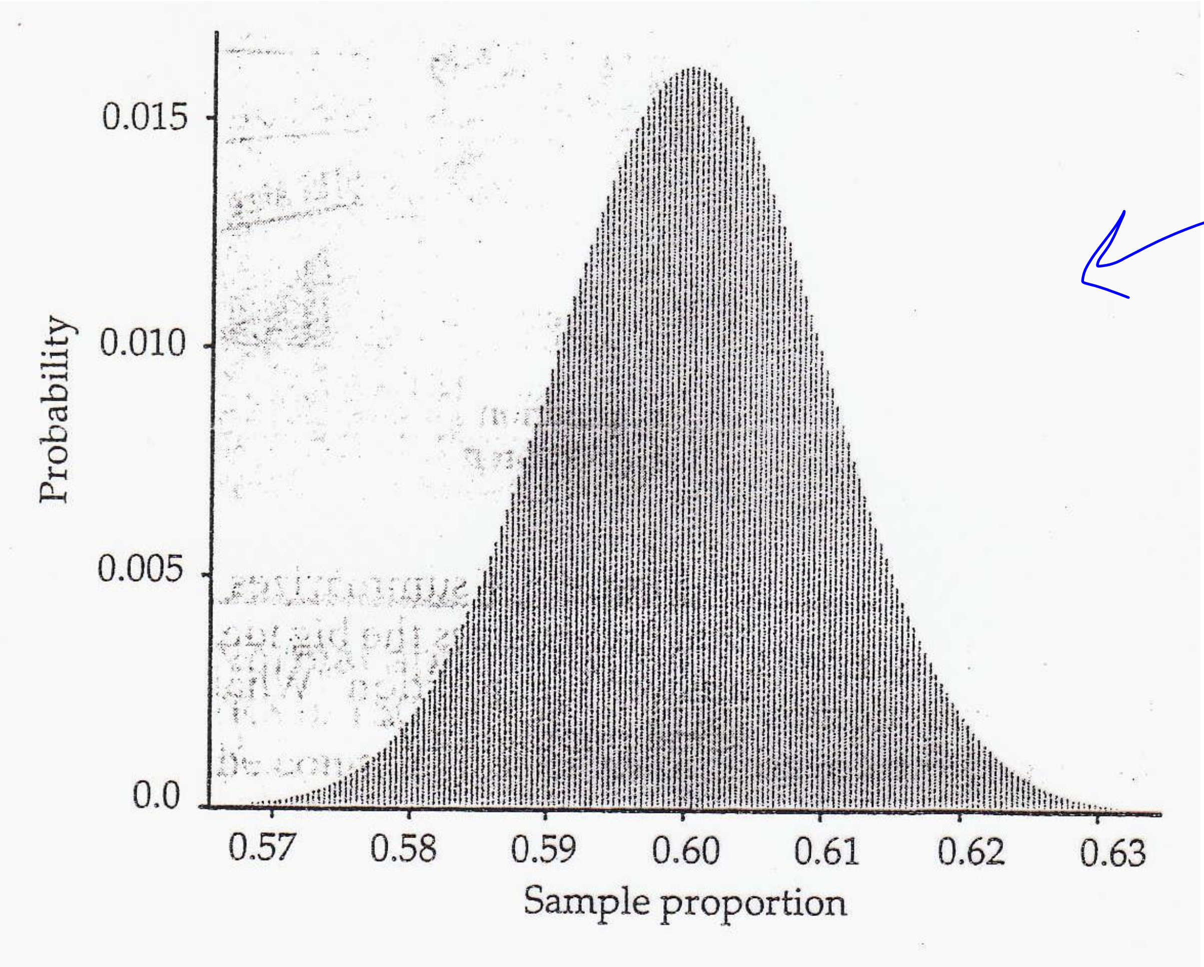
$\mu_{\hat{p}} = p$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

$= \sqrt{\frac{pq}{n}}$

Example: A sample survey asks a nationwide random sample of 2500 adults if they agree or disagree that «I like buying new clothes, but shopping is often frustrating and time-consuming». Suppose that 60% of all adults would agree if asked this question.

Figure below is a probability histogram of the exact distribution of the proportion of frustrated shoppers $\hat{p}$, based on the binomial distribution with $n = 2500$, $p = 0.6$.



$$\hat{p} = \frac{X}{n}$$

The histogram looks very Normal.

Normal Approximation for Counts and Proportions:

Draw an SRS of size n from a large population having population proportion p of successes.

Let X be the count of successes in the sample and $\hat{p} = X/n$ be the sample proportion of successes.

When n is large, the sampling distributions of these statistics are approximately Normal:

$$X \text{ is approximately } N(np, \sqrt{np(1-p)})$$

$$\hat{p} \text{ is approximately } N(p, \sqrt{\tfrac{p(1-p)}{n}})$$

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Example: Let's compare the Normal approximation with the exact calculation.

$X \sim Bin(2500, 0.6)$, $P(\hat{p} \geq 0.58) = 0.9802$ (software)

$P(\hat{p} \geq 0.58)$

$np = 2500 \cdot 0.6$
$= 1500 > 10$

$n(1-p) = 2500 \cdot 0.4$
$= 1000 > 10$

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

$$= N\left(0.6, \sqrt{\frac{0.6 \cdot 0.4}{2500}}\right) = N(0.6, 0.0098)$$

$$P(\hat{p} \geq 0.58) = P(Z \geq -2.04)$$

z-score for 0.58 $= \dfrac{0.58 - 0.6}{0.0098} = -2.04$

$$= 1 - P(Z \leq -2.04) = 1 - 0.0207$$

$= 0.9793$ close

Software $= 0.9802$

Example: The audit described in the example from the previous lecture examined an SRS of 150 sales records for compliance with sales tax laws. In fact, 8% of all the company's sales records have an incorrect sales tax classification. The count X of bad records in the sample has approximately the Bin(150, 0.08) distribution.

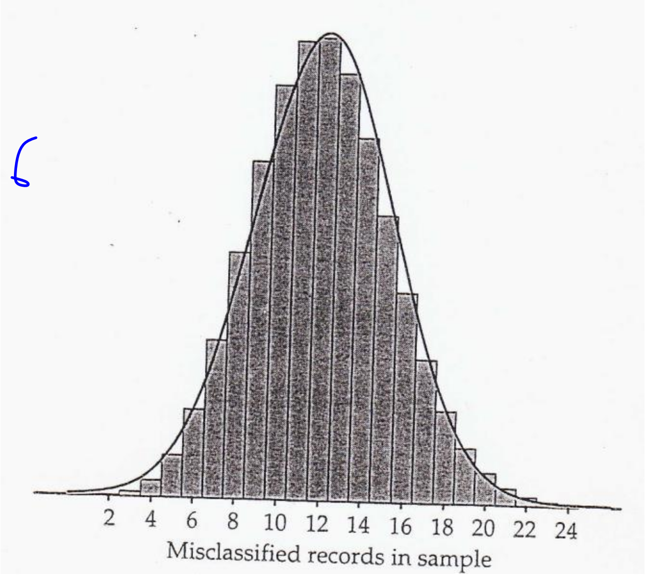
$$\boxed{P(X \leq 10) = 0.3384 \text{ (software)}}$$

$$X \sim N(np, \sqrt{np(1-p)}) = N(150 \cdot 0.08, \sqrt{150 \cdot 0.08 \cdot 0.92})$$

$$= N(12, 3.3226)$$

$$P(X \leq 10) = P(Z \leq -0.6)$$

$$\text{z-score for } 10 = \frac{10-12}{3.3226} = -0.6$$



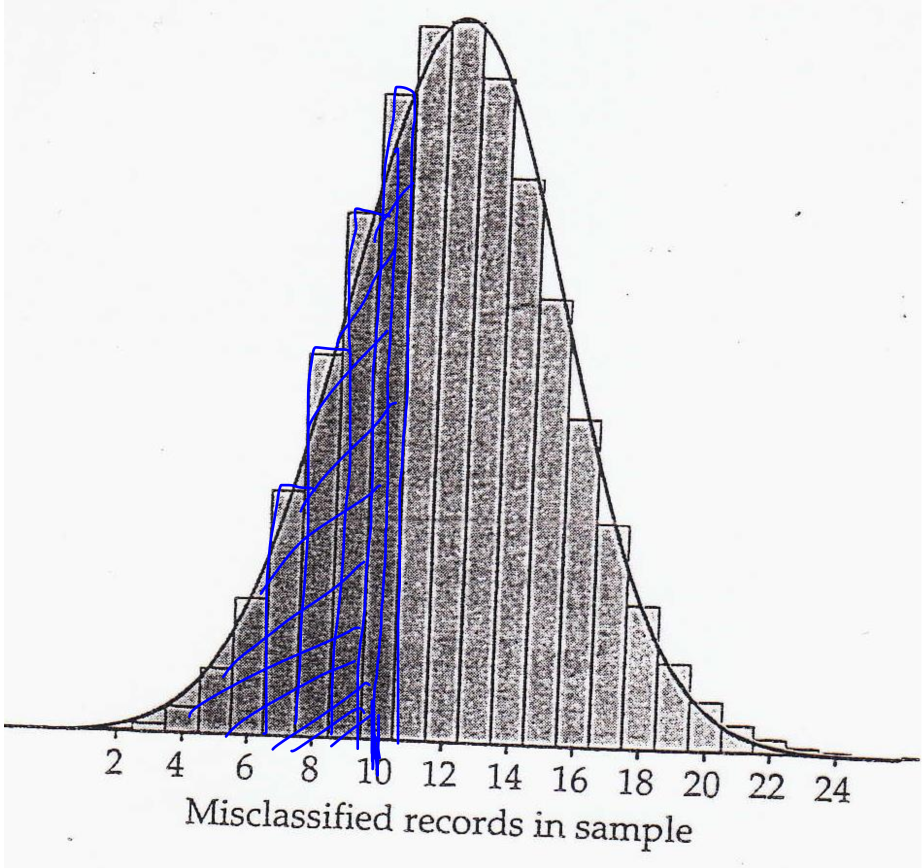
$=$ table 0.2743 not that close

0.3384

Why? $np = 12 \rightarrow$ close to 10

Continuity Correction

Figure below illustrates an idea that greatly improves the accuracy of the Normal approximation to binomial probabilities.



$P(X \leq 10) = P(X \leq 10.5)$

Z-score for 10.5

$= \frac{10.5 - 12}{3.3226}$

$= -0.45$

$= P(Z \leq -0.45) = 0.3264$

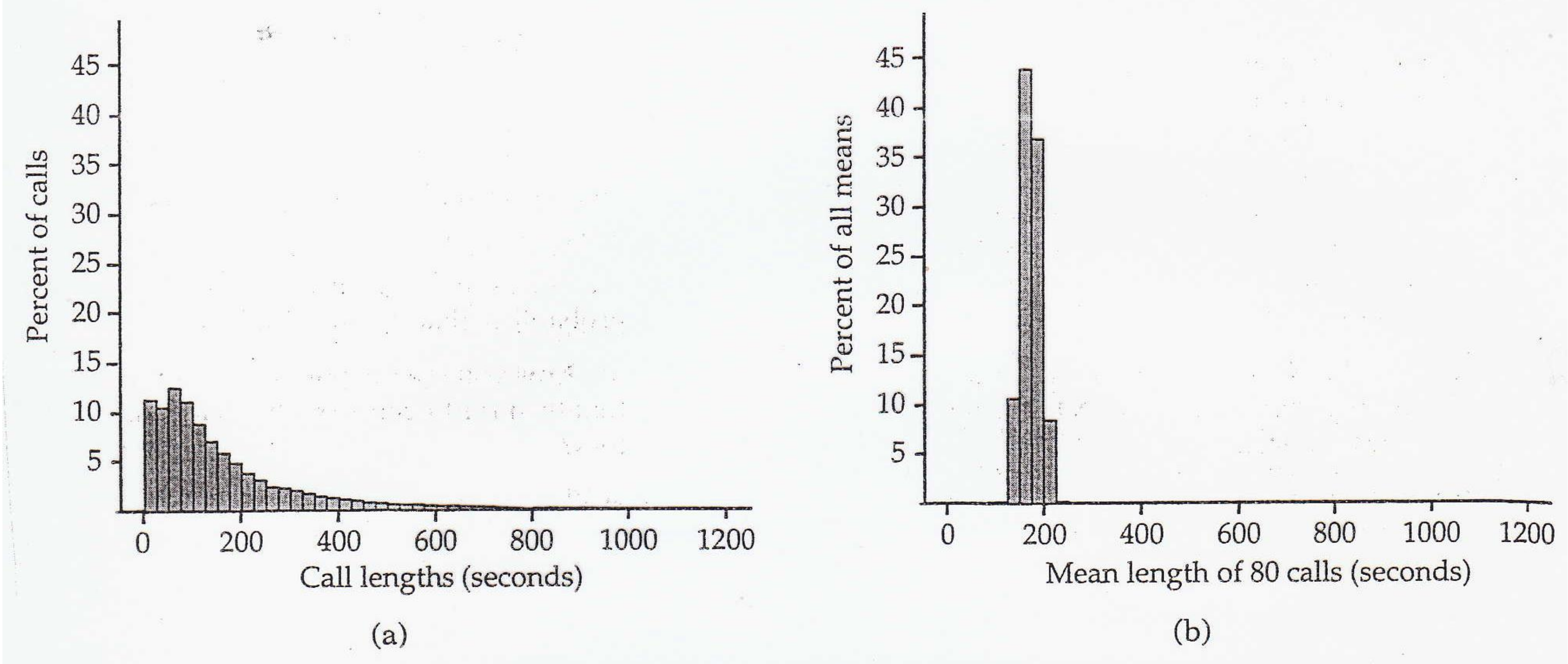
software $= 0.3384$

[illegible]

Sampling Distribution of Sample Mean

Sample means are among the most common statistics, and we are often interested in their sampling distribution.

The figure below shows (a) the distribution of lengths of all customer service calls received by a bank in a month ($\mu = 170$); (b) the distribution of the sample means $\bar{x}$ for 500 random samples of size 80 from this population.



Facts about sample means:

- Sample means are less variable than individual observations.
- Sample means are more Normal than individual observations.

Mean and Standard Deviation of $\bar{X}$

The sample mean $\bar{x}$ from a sample or an experiment is an estimate of the mean μ of the underlying population.

Let $X_1, X_2, \dots, X_n$ be taken from a population with mean μ and standard deviation σ.

We say, $X_1, X_2, \dots, X_n \sim$ i.i.d. with mean μ and st. deviation σ.

i.i.d. = independent identically distributed

$$\mu_{X_i} = \mu$$

$$\sigma_{X_i} = \sigma$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$= \frac{X_1 + X_2 + \dots + X_n}{n}$$

Then

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Why?

$$\mu_{\bar{X}} = \mu_{\frac{1}{n}\Sigma X_i} = \frac{1}{n}\mu_{\Sigma X_i} = \frac{1}{n}\mu_{X_1+\dots+X_n}$$

$$= \frac{1}{n}\left[\mu_{X_1} + \dots + \mu_{X_n}\right] = \frac{1}{n}\left[\underbrace{\mu + \dots + \mu}_{n \text{ times}}\right]$$

$$= \frac{1}{\cancel{n}} \cdot \cancel{n}\mu = \mu$$

$$\sigma^2_{\bar{X}} = \sigma^2_{\frac{1}{n}\Sigma X_i} = \frac{1}{n^2}\sigma^2_{X_1+\dots+X_n}$$

$$= \frac{1}{n^2}\left[\sigma^2_{X_1} + \dots + \sigma^2_{X_n}\right] = \frac{1}{n^2}\left[\underbrace{\sigma^2 + \dots + \sigma^2}_{n \text{ times}}\right]$$

$$= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem (CLT)

We have described the center and spread of the probability distribution of $\bar{X}$. What about its shape?

It can be shown that if we have a population from

$$N(\mu, \sigma)$$

then the distribution of sample mean of n independent observations is

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

Moreover,

<u>CLT</u>: Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large enough, the sampling distribution of the sample mean is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Example: Let's look at the histogram of the lengths of telephone calls again. For that example, $\underbrace{\mu = 170}$ and $\sigma = \underbrace{184.81}$ seconds. Consider a sample of size 80.

$$\mu_{\bar{X}} = \mu = 170$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{184.81}{\sqrt{80}} = 20.66$$

How close will the sample mean be to the population mean? $\bar{X} \underset{\overset{\downarrow}{\text{CLT}}}{\sim} N(170, 20.66)$

By 68-95-99.7% rule,

95% of all samples will have $\bar{X}$ within $2\sigma_{\bar{X}} = 2 \cdot 20.66 = 41.32$ of μ

$$\bar{X} \in (\mu - 41.32, \mu + 41.32)$$

How can we reduce the standard deviation?

$$\sigma_{\bar{x}} = \frac{\sigma \rightarrow \text{fixed}}{\sqrt{n} \searrow \text{increase}}$$

To reduce $\sigma_{\bar{x}}$ by 4,

take $n = 80 \cdot 16 = 1280$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{134.81}{\sqrt{1280}} = 5.165$$

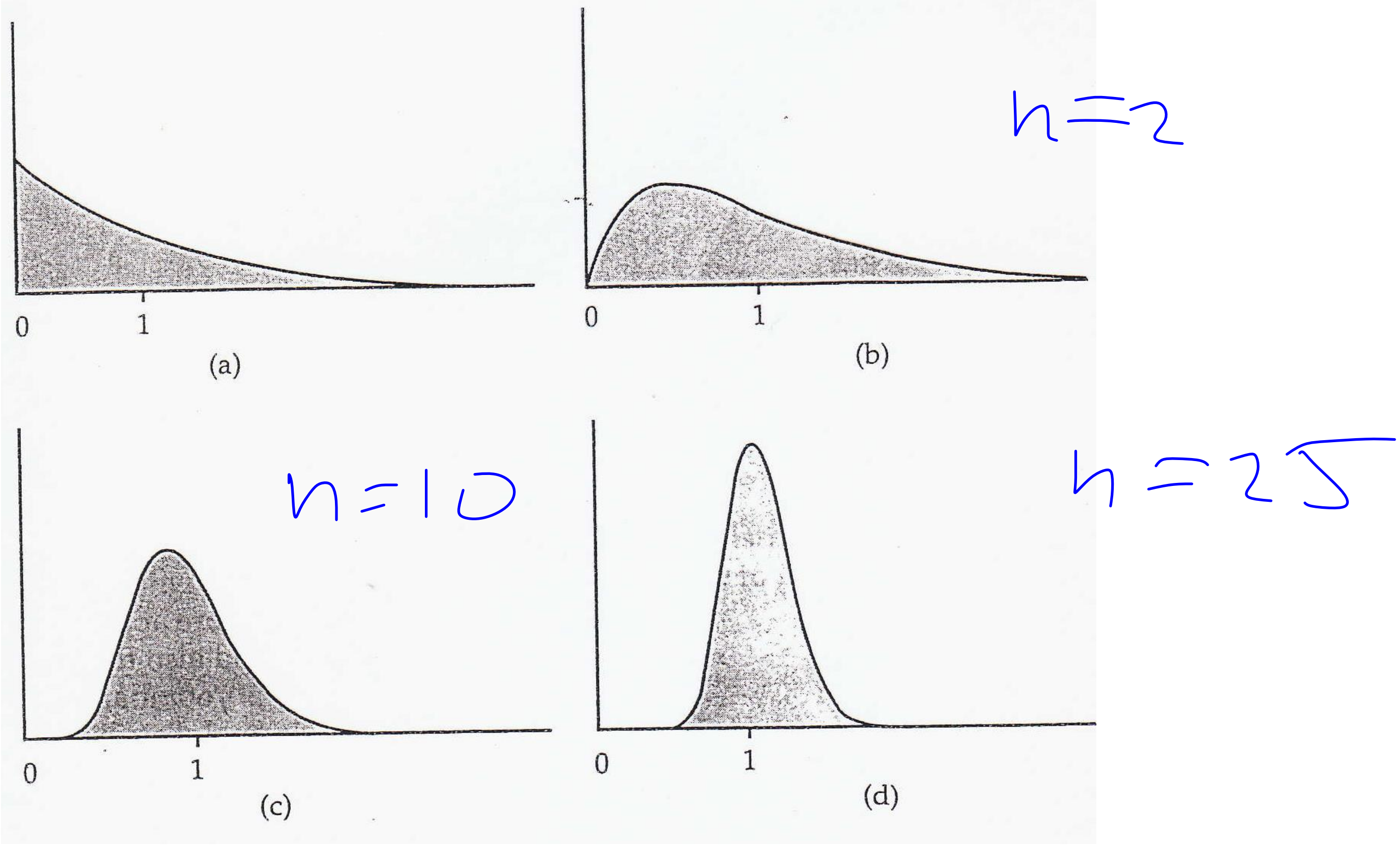
$$2 \cdot \sigma_{\bar{x}} = 10.33$$

$$\bar{x} \in (\mu - 10.33, \mu + 10.33)$$

95% of the time

Figure below shows the CLT in action for another very non-Normal population: (a) displays the density curve of a single observation, that is, of the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0. The mean of this distribution is 1, and its standard deviation is also 1. This particular continuous distribution is called an **exponential distribution**.

Figures (b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population.



Example: The time X that a technician requires to perform preventive maintenance on an air-conditioning unit is governed by the exponential distribution.

The mean time is $\boxed{\mu = 1}$ hour and the standard deviation is $\boxed{\sigma = 1}$ hour. Your company operates 70 of these units. What is the probability that their average maintenance time exceeds 50 minutes?

Let $\bar{X}$ = sample mean time spent working on 70 units. $n = 70$

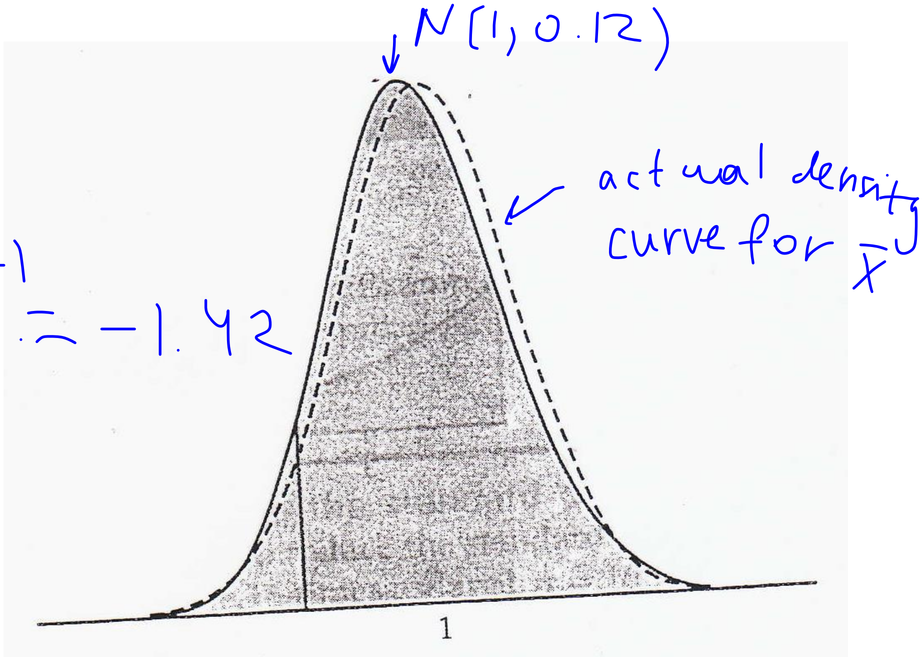
$$P(\bar{X} > 50 \text{ min}) = ?$$

$$\bar{X} \underset{CLT}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(1, \frac{1}{\sqrt{70}}\right)$$

$$= N(1, 0.12)$$

$$50 \text{ min} = \frac{50}{60} \text{ hour} = 0.83 \text{ h}$$

$$P(\bar{X} > 0.83)$$



z-score for 0.83 $= \dfrac{0.83-1}{0.12} = -1.42$

$$= P(Z > -1.42)$$

$$= 0.9222$$

table

close

Actual probability = 0.9294

A few more facts:

- The Normal approximation for sample proportions and counts is an example of the CLT.
  Why?

$\hat{p} = \frac{X}{n}$, $X = \sum X_i$, X_i are Bernoulli

$\mu_{X_i} = p$, $\sigma_{X_i} = \sqrt{p(1-p)}$

$$\hat{p} = \frac{1}{n} \sum X_i \underset{CLT}{\sim} N\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

- Any linear combination of independent Normal random variables is also Normally distributed.

$X \sim N(\mu_x, \sigma_x)$, $Y \sim N(\mu_y, \sigma_y)$

$$aX + bY \sim N\left(a\mu_x + b\mu_y, \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}\right)$$

Example: Tom and George are playing in the club golf tournament.

Their scores vary as they play the course repeatedly.

Tom's score X has $N(110, 10)$ distribution

George's score Y varies from round to round according to $N(100, 8)$ distribution.

If they play independently, what is the probability that Tom will score lower than George and thus do better in the tournament?

$$? = P(X < Y) = P(X - Y < 0)$$

$$\mu_{X-Y} = \mu_X - \mu_Y = 110 - 100 = 10$$

$$\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2} = \sqrt{10^2 + 8^2} = \sqrt{164} = 12.8$$

So, $X - Y \sim N(10, 12.8)$

$P(X - Y < 0) =$

z-score for 0 $= \dfrac{0-10}{12.8} = -0.78$

$= P(Z < -0.78)$

$= 0.2177$