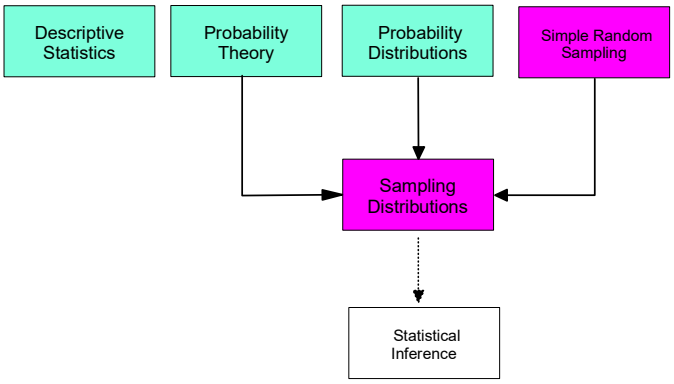
The Course So Far:



- **Simple random sampling:** A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained.
- **Simple random sample:** A sample obtained by simple random sampling.

Governor (G)
Lt. Governor (L)
Secretary of State (S)
Attorney General (A)
Treasurer (T)

$$\binom{5}{2} = 10$$

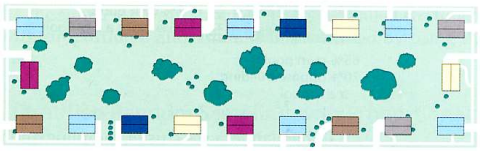
22	166	310	454	598
70	214	358	502	646
118	262	406	550	694

Systematic Random Sampling

Step 1 Divide the population size by the sample size and round the result down to the nearest whole number, m.

Step 2 Use a random-number table (or a similar device) to obtain a number, k, between 1 and m.

Step 3 Select for the sample those members of the population that are numbered k, k + m, k + 2m,....



Cluster Sampling

Step 1 Divide the population into, groups (clusters).

Step 2 Obtain a simple random sample of the clusters.

Step 3 Use all the members of the clusters obtained in Step 2 as the sample.

Stratified Random Sampling with Proportional Allocation

Step 1 Divide the population into subpopulations (strata).

Step 2 From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.

Step 3 Use all the members obtained in Step 2, as the sample.

Definition: Sampling Error

Sampling Error is the error resulting from using a sample to estimate a population characteristic

Definition: Sampling Distribution of the Sample Mean

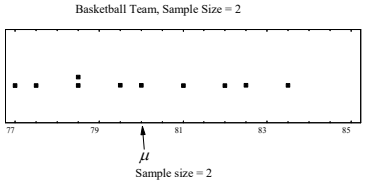
For a variable *x* and a given sample size, the distribution of the $\bar{x}$ variable -- that is, the distribution of all possible sample means -- is called the *sampling distribution of the sample mean.*

In statistics, the following terms and phrases are synonymous.
- Sampling distribution of the sample mean
- Distribution of the variable $\bar{x}$
- Distribution of all possible sample means of a given sample size

Player	A	B	C	D	E
Height	76	78	79	81	86

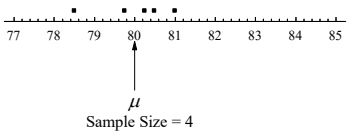
$$\mu = \frac{\sum x}{N} = \frac{76+78+79+81+86}{5} = 80 \text{ inches.}$$

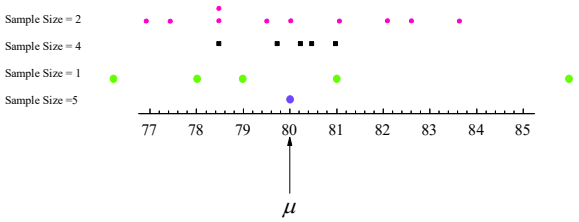
Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5



Sample Size = 4

Sample	Heights	$\bar{x}$
A, B, C, D	76, 78, 79, 81	78.50
A, B, C, E	76, 78, 79, 86	79.75
A, B, D, E	76, 78, 81, 86	80.25
A, C, D, E	76, 79, 81, 86	80.50
B, C, D, E	78, 79, 81, 86	81.00





$$\bar{x} = \frac{\sum x}{\text{k}} \quad \frac{\text{all x in sample}}{\text{k}} \text{ where k is the sample size.}$$

$$E[\bar{x}] = \mu_{\bar{x}} = \int_{\text{all values of } \bar{x}} \bar{x} f(\bar{x}) d\bar{x} = \sum_{\text{all values of } \bar{x}} \bar{x} P(\bar{x})$$

Continuous Distribution — Discrete Distribution

Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5

Or, for our 5-man basketball team sampling two at a time:

$$E[\bar{x}] = \mu_{\bar{x}} = \int_1^{10} \bar{x}\left(\frac{1}{10}\right) d\bar{x} = \frac{\int_1^{10} \bar{x} d\bar{x}}{10} = \frac{\sum_{i=1}^{10} \bar{x}_i}{10} = 80$$

$$= \frac{77+77.5+78.5+81+78.5+79.5+79.5+82+80+82.5+83.5}{10} = 80.0$$

or, more generally:

since

$$\bar{x} = \frac{\sum x}{\text{k}} \quad \frac{\text{all x in sample}}{\text{k}} \text{ where k is the sample size.}$$

using the algebra of expectations for sample size = *k*:

$$E[\bar{x}] = \mu_{\bar{x}} = E\left[\frac{1}{k}\sum_{i=1}^{k} x_i\right] = \frac{1}{k} E\left[\sum_{i=1}^{k} x_i\right] = \frac{1}{k}\sum_{i=1}^{k} E[x_i] = \frac{1}{k} \bullet k\mu = \mu.$$

$$E[\bar{x}] = \mu$$

The General Rule for the mean (expected value) of $\bar{x}$:

For samples of size *k*, the mean (expected value) of the variable $\bar{x}$ equals the mean of the variable under consideration:

$$\mu_{\bar{x}} = \mu.$$

In other words, for each sample size, the mean of all possible sample means equals the population mean.

The Variance of the Sample Mean:

Let $x_1, x_2, \ldots x_k$ be independent random variables, each having the same known variance $V[x_i] = \sigma_x^2$ then,

$$V[\bar{x}] = V\left[\frac{1}{k}(x_1 + x_2 + \cdots + x_k)\right] \text{ by definition}$$

$$= \left(\frac{1}{k}\right)^2 V[x_1 + x_2 + \cdots + x_k] \text{ (expectation rule)}$$

$$= \left(\frac{1}{k}\right)^2 \{V[x_1] + V[x_2] + \cdots + V[x_k]\} \text{ (another expectation rule)}$$

$$\sigma_{\bar{x}}^2 = \left(\frac{1}{k}\right)^2 k\sigma_x^2 = \frac{1}{k}\sigma_x^2 \text{ since } V[\tilde{x}_i] = \sigma_x^2.$$

or $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{k}}$. The Standard Deviation of the Sample Mean

The Standard Deviation of $\bar{x}$

For samples of size *k*, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{k}}.$$

In other words, for each sample size, the standard deviation of all possible sample means equals the population standard deviation divided by the square root of the sample size, *assuming that we know the population standard deviation.*

SAMPLE SIZE AND SAMPLING ERROR (REVISITED)

The possible sample means cluster more closely around the population mean as the sample size increases, and therefore the larger the sample size, the smaller the sampling error tends to be in estimating a population mean by a sample mean. Here is why that key fact is true.

• The larger the sample size, the smaller is the standard deviation of $\bar{x}$

• The smaller the standard deviation of $\bar{x}$ the more closely the possible values of $\bar{x}$ (the possible sample means) cluster around the mean of $\bar{x}$.

• The mean of $\bar{x}$ equals the population mean.

Because the standard deviation of $\bar{x}$ determines the amount of sampling error to be expected when a population mean is estimated by a sample mean, it is often referred to as **the standard error of the sample mean.** In general, the standard deviation of a statistic used to estimate a parameter is called the **standard error (SE)** of the statistic.

The Distribution of $\bar{x}$ (normal population) $= N\left(\mu, \frac{\sigma^2}{n}\right)$.

That is, the sampling distribution of sample means drawn from a normally distributed population is itself normal.

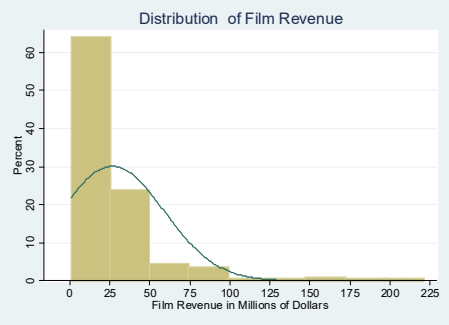
The Central Limit Theorem:

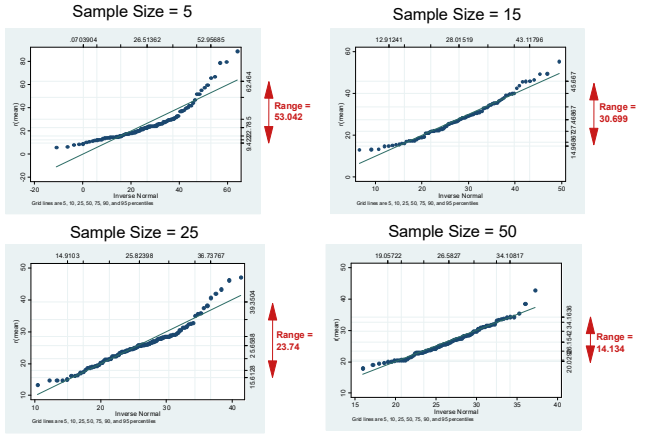
If random samples of *n* observations are drawn from an arbitary population with finite mean μ and standard deviation σ, then when *n* is large, the sampling distribution of the sample mean is ***approximately*** **normally distributed with mean and standard deviation**

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

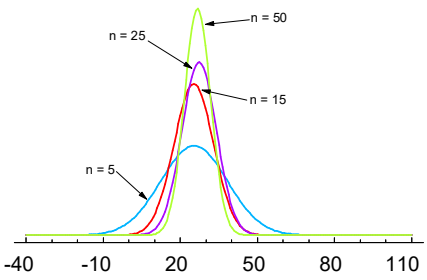
The approximation will become more and more accurate as *n* becomes larger and larger.

Film Revenue Data for Central Limit Theorem Exercise





Sampling Distributions for Sample Mean: n=5, 15, 25, 50



Sample Means of Film Revenue

Recap: The Difference between a Sampling Distribution and a Probability Distribution:

- A **probability distribution of a random variable** gives, for each value of the random variable, the probability that a *given value* of the random variable will be obtained.
- The **sampling distribution of a sample statistic** shows, for a sample statistic computed from a given sample size, the probability distribution of *all possible values* of the sample statistic that can be computed from samples of that size.
- So, for a variable *x* and a given sample size *n*, the distribution of the variable $\bar{x}$, that is, the distribution of all possible sample means -- is called the **sampling distribution of the sample mean**.

Finite Population Correction Factor for the Standard Error of the Mean

Finite Population Multiplier: $\sqrt{(N-n)/(N-1)}$

Variance of $\bar{x}$: $$\sigma_{\bar{x}}^2 = \frac{1}{n}\sigma^2\left(\frac{N-n}{N-1}\right)$$

Standard error of the mean: $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{\text{N - n}}{\text{N - 1}}}$$

The Sample Variance & Standard Deviation

$$s^2 = \left(\frac{1}{n-1}\right)\bullet\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

The sample standard deviation is simply:

$$s = \sqrt{s^2}$$

Proof that the Sample Variance is an Unbiased Estimator of the Population Variance

Early in the course I claimed that the "best" estimator of the population variance, *sigma-squared* is *s-squared* defined as:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

even though it would seem that a better estimator would be: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

So, let $x_1, x_2, \ldots, x_n$ be a random sample with $E(x_i) = \mu$ and $V(x_i) = \sigma^2$. Show that s^2 is a biased estimator for σ^2 and s^2 is an unbiased estimator for σ^2.

First, with some basic algebra (which I'll leave to you) we can demonstrate that:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2.$$

Then, we can write the *expected value* of this sum of squared differences as:

$$E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = E\left(\sum_{i=1}^{n}x_i^2\right) - nE\left(\bar{x}^2\right) = \sum_{i=1}^{n}E(x_i^2) - nE\left(\bar{x}^2\right).$$

The expected value of a difference is the difference of the expected values. And the expected value of a sum is the sum of the expected values.

Notice that $E\left(\sum x_i^2\right)$ is the same for $i = 1, 2, \ldots, n$. We use this and the fact that the variance of a random variable is given by $V(x) = E(x^2) - [E(x)]^2$ to conclude that

$$E(x_i^2) = V(x_i) + [E(x_i)]^2 = \sigma^2 + \mu^2, E(\bar{x}^2) = V(\bar{x}) + [E(\bar{x})]^2 = \sigma^2/n + \mu^2,$$

and that

$$E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

It follows that

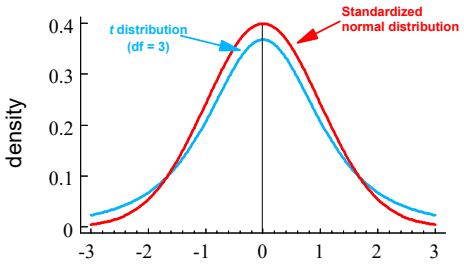
$$E(s^2) = \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n}(n-1)\sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2$$

and that s^2 is biased because $E(s^2) \neq \sigma^2$. However,

$$E(s^2) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

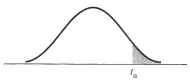
so we see that s^2 is an unbiased estimator for σ^2.

Comparing the t- and normal distributions



z- and t-values

d.f.	t.100	t.050	t.025	t.010	t.005	d.f.
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.



Critical Values of *t*
for
Student's *t* distribution

d.f.	t.100	t.050	t.025	t.010	t.005	d.f.
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

Critical Values of *t*
for
Student's *t* distribution

$\bar{x} = 126.8$

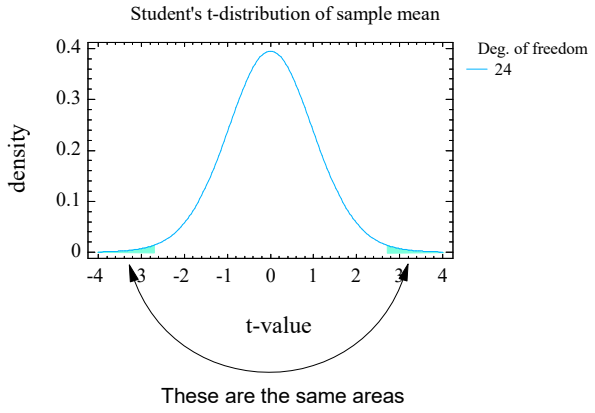
$s = 6.0$

$P(\bar{x} \leq 126.8)$ if $\mu = 130 = ?$

$$P(\bar{x} \leq 126.8) = P\left(\frac{\bar{x} - \mu}{s/\sqrt{n}} \leq \frac{126.8 - 130}{6/\sqrt{25}}\right)$$
$$= P\left(t \leq \frac{-3.2}{1.2}\right) = P(t \leq -2.667)$$

$df = 25 - 1 = 24.$

$0.01 > P(t \leq -2.667) > 0.005$



These are the same areas

Using the normal in place of t-distribution

- When n is large (>30) $\bar{x}$ will *at a minimum* be approximately normally distributed.
- When *n* is large *s* will usually be a good approximation to sigma.
- In that case, the distribution of $t=(\bar{x}-\mu)/(s/\sqrt{n})$ and that of $z=(\bar{x}-\mu)/(\sigma/\sqrt{n})$ will be approximately the same.
- So, for large samples we can use the standard normal distribution to approximate the t-distribution.

Sampling Distribution of the Sample Proportion

$\hat{p}=\frac{x}{n}$ (the sample proportion)

$$E[\hat{p}]=\mu_{\hat{p}}=E\left[\frac{x}{n}\right]=\frac{1}{n}E[x]=\frac{1}{n}\bullet np=p$$

$$V[\hat{p}]=\sigma_{\hat{p}}^2=V\left[\frac{x}{n}\right]=\frac{1}{n^2}V[x]=\frac{np(1-p)}{n^2}=\frac{p(1-p)}{n}$$

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

$$\hat{p}=\frac{x}{n}=\frac{8}{300}=0.02667$$

$$\sigma_{\hat{p}}=\sqrt{\frac{p(1-p))}{n}}\approx\sqrt{\frac{\hat{p}(1-\hat{p}))}{n}}=0.00930$$

$$z=\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p}))}{n}}}=\frac{0.02667-0.02}{0.00930}=0.71720$$

$p=0.02$