Statistical Models and Sampling Distributions

Joint Distribution of Observations (Statistical Models for Data)

We regard $x_1, \ldots, x_n$ as realizations or observations of n random variables $X_1, \ldots, X_n$. The joint distribution of $X_1, \ldots, X_n$ describes the probability distribution of different values of $x_1, \ldots, x_n$ under a prescribed sampling scheme.

Random Samples and Population Distribution

Definition: The random variables $X_1, X_2, \ldots, X_n$ are said to be a random sample of size n from a population distribution $p(x)$ if

1. $X_1, X_2, \ldots, X_n$ are *independent* random variables.
2. $X_1, X_2, \ldots, X_n$ have identical probability distribution, $p(x)$.

This model is used to model n observations $x_1, \ldots, x_n$ drawn with simple random sampling from a population. $p(x)$ is determined by the population, so called **population distribution**.

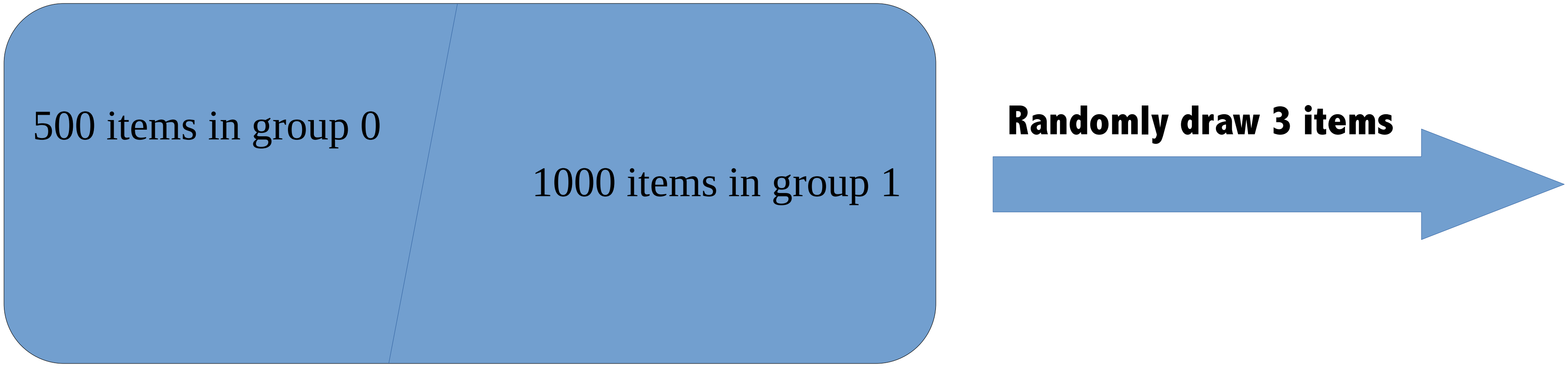
In such cases, the joint distribution of $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

Sometimes, we also say that $X_1, \ldots, X_n$ are independently and identically distributed (shortened by I.I.D. or i.i.d.) with $p(x)$:

$$X_1, \ldots, X_n \overset{\text{I.I.D.}}{\sim} p(x)$$

Example 1:



500 items in group 0

1000 items in group 1

Randomly draw 3 items

$$
\begin{aligned}
P(X_1 = 1, X_2 = 0, X_3 = 1) &= 2/3 \times 1/3 \times 2/3 \\
&= (2/3)^1 (1/3)^0 \times (2/3)^0 (1/3)^1 \times (2/3)^1 \times (1/3)^0
\end{aligned}
$$

Let $p(x) = P(X = x) = (2/3)^x (1 - 2/3)^{1-x}$, then we can write $P(X_1 = 1, X_2 = 0, X_3 = 1) = p(1)p(0)p(1)$.

Now, let's become more abstract. Suppose $x_1, \ldots, x_n$ be a particular observations of $X_1, \ldots, X_n$, then the joint distribution of $X_1, \ldots, X_n$ is written as:

$$
\begin{aligned}
f(x_1, \ldots, x_n) &= \prod_{i=1}^{n} p(x_i) \\
&= \prod_{i=1}^{n} (2/3)^{x_i} (1/3)^{1-x_i} \\
&= (2/3)^{\sum_{i=1}^{n} x_i} (1 - 2/3)^{n - \sum_{i=1}^{n} x_i}
\end{aligned}
$$

If the proportion of items with label 1, θ, is unknown, the probability model for $x_1, \ldots, x_n$ is given by

$$
\begin{aligned}
f(x_1, \ldots, x_n) &= \prod_{i=1}^{n} p(x_i) \\
&= \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i} \\
&= \theta^{x} (1 - \theta)^{n-x}
\end{aligned}
$$

where $x = \sum_{i=1}^{n} x_i$.

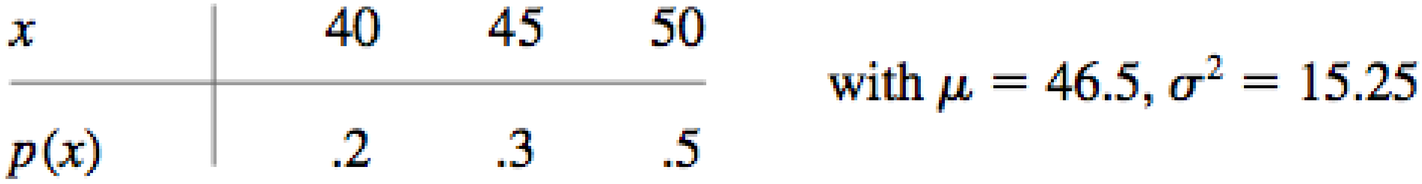
For this example, we say that binary random variables $X_1, \ldots, X_n$ are a random sample (I.I.D.) from the population distribution Bernoulli (θ):

$$X_1, \ldots, X_n \overset{\text{I.I.D.}}{\sim} \text{Bernoulli}(\theta)$$

The population distribution is $p(x) = \theta^x (1-\theta)^{1-x}$ for $x = 0 \text{ or } 1$.

Example 2:

A large automobile service center charges \$40, \$45, and \$50 for a tune-up of four-, six-, and eight-cylinder cars, respectively. If 20% of its tune-ups are done on four-cylinder cars, 30% on six-cylinder cars, and 50% on eight-cylinder cars. For this example, the population distribution of revenue x from a single randomly selected tune-up is given.



x	40	45	50
$p(x)$	.2	.3	.5

with $\mu = 46.5$, $\sigma^2 = 15.25$

The revenues of n randomly selected tune-ups, denoted by $x_1, \ldots, x_n$, are said a random sample of size *n* from $p(x)$.

Example 3:

Suppose $X_1, \ldots, X_n$ are service times at the cash register in a mini market for *n* different customers. Let's **assume** that the time to serve a customer, X_i, has an exponential distribution with parameter λ. We say that $X_1, \ldots, X_n$ is a random sample from exponential distribution with parameter λ. The joint probability density (probability model) of $X_1, \ldots, X_n$ is given by

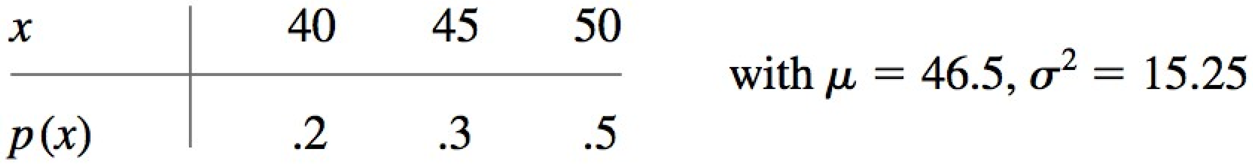
$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i)$$

Sampling Distributions of Statistic

Definition: *A statistic is a function of sample data*. Prior to obtaining data, there is variability as to what value of any particular values that sample data will result. Therefore, sample data is regarded as random variables. Accordingly, a statistic is a function of random variables. Therefore, a statistic is also a random variable. Conventionally, a statistic will be denoted by an uppercase letter when it is a random variable; a lowercase letter is used to represent the calculated or observed value of the statistic.

Example:

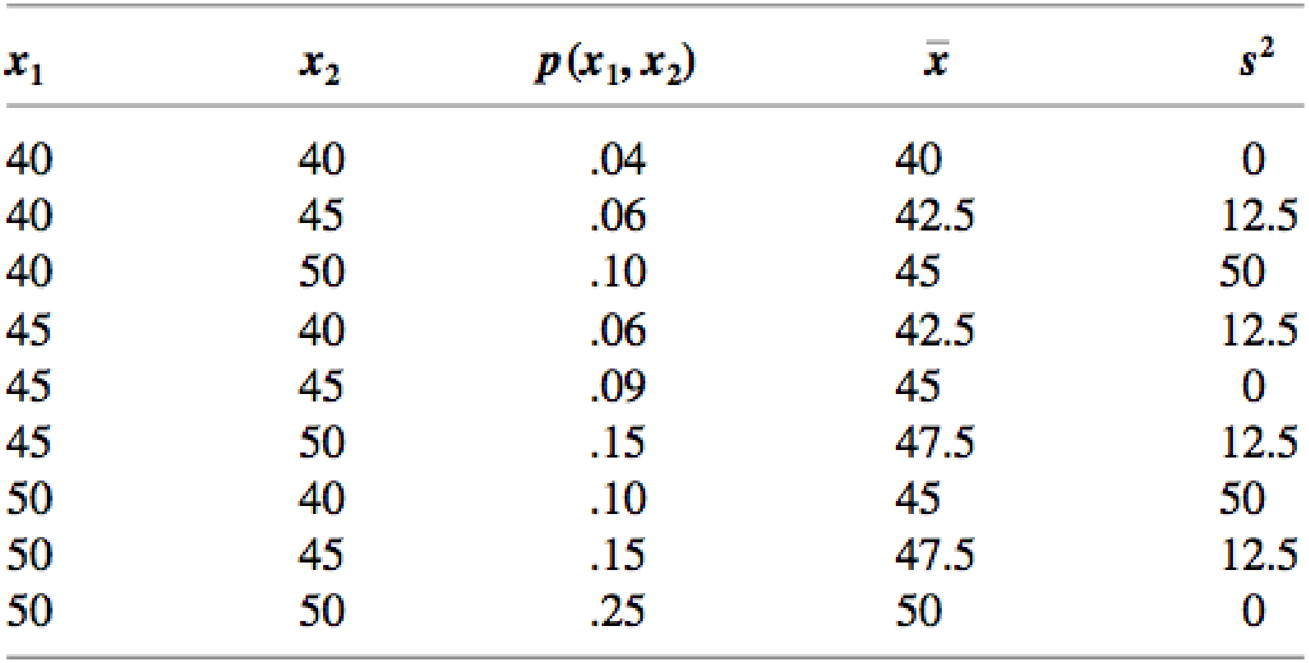
A large automobile service center charges \$40, \$45, and \$50 for a tune-up of four-, six-, and eight-cylinder cars, respectively. If 20% of its tune-ups are done on four-cylinder cars, 30% on six-cylinder cars, and 50% on eight-cylinder cars. For this example, the population distribution of revenue x from a single randomly selected tune-up is given:



x	40	45	50
$p(x)$	.2	.3	.5

with $\mu = 46.5$, $\sigma^2 = 15.25$

The joint distribution of two observations x_1 and x_2, and $\bar{x}$ and s^2:



x_1	x_2	$p(x_1, x_2)$	$\bar{x}$	s^2
40	40	.04	40	0
40	45	.06	42.5	12.5
40	50	.10	45	50
45	40	.06	42.5	12.5
45	45	.09	45	0
45	50	.15	47.5	12.5
50	40	.10	45	50
50	45	.15	47.5	12.5
50	50	.25	50	0

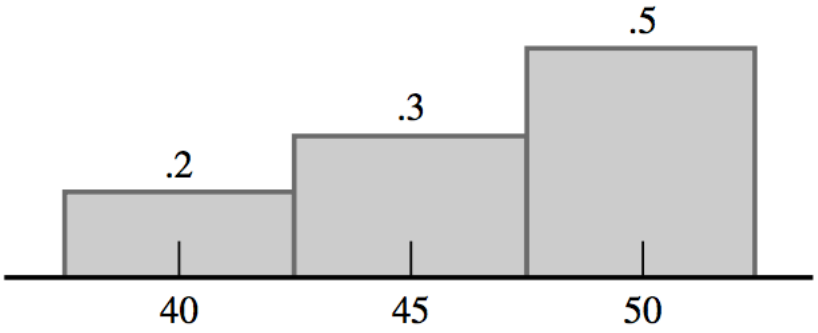
The sampling distribution of $\bar{x}$ and s^2 based on 2 observations:

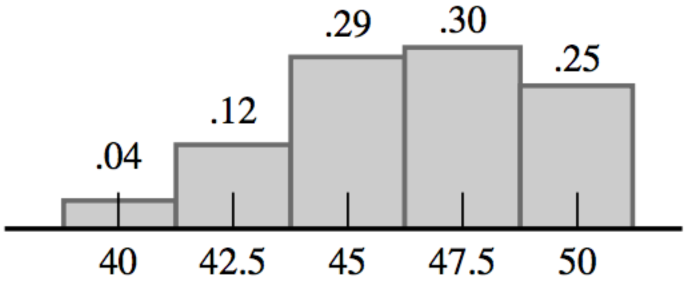
$$p_{\bar{X}}(45) = P(\bar{X} = 45) = .10 + .09 + .10 = .29$$

$$\begin{aligned} p_{s^2}(50) &= P(S^2 = 50) = P(X_1 = 40, X_2 = 50 \quad \text{or} \quad X_1 = 50, X_2 = 40) \\ &= .10 + .10 = .20 \end{aligned}$$

$\bar{x}$	40	42.5	45	47.5	50
$p_{\bar{X}}(\bar{x})$	.04	.12	.29	.30	.25

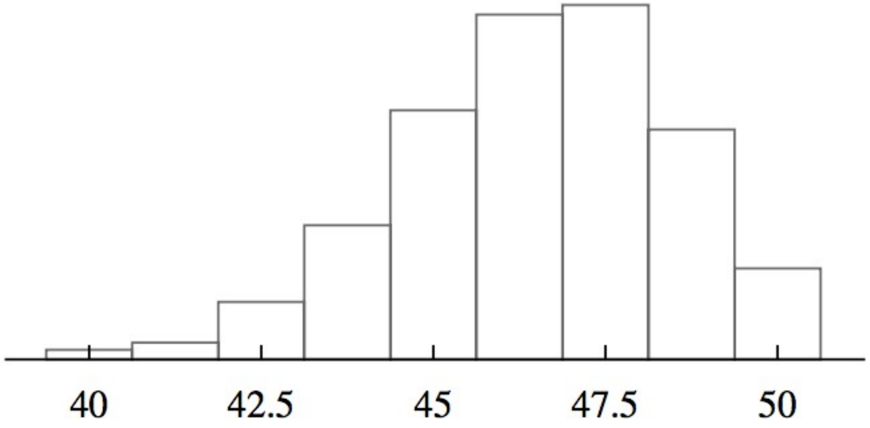
s^2	0	12.5	50
$p_{S^2}(s^2)$	.38	.42	.20





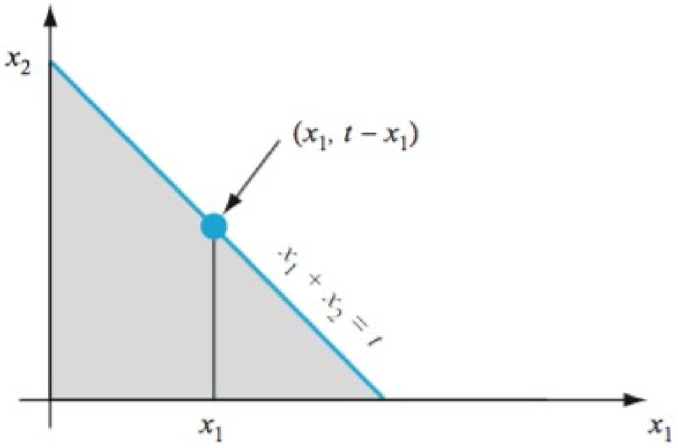
The sampling distribution of $\bar{x}$ based on 4 observations:

$\bar{x}$	40	41.25	42.5	43.75	45	46.25	47.5	48.75	50
$p_{\bar{X}}(\bar{x})$	.0016	.0096	.0376	.0936	.1761	.2340	.2350	.1500	.0625



Example:

Suppose X_1, X_2 are IID with $\exp(\lambda)$. $T_0 = X_1 + X_2$



$$F_{T_o}(t) = P(X_1 + X_2 \leq t) = \iint\limits_{\{(x_1, x_2): x_1 + x_2 \leq t\}} f(x_1, x_2)\, dx_1\, dx_2$$

$$= \int_0^t \int_0^{t - x_1} \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2}\, dx_2\, dx_1 = \int_0^t (\lambda e^{-\lambda x_1} - \lambda e^{-\lambda t})\, dx_1$$

$$= 1 - e^{-\lambda t} - \lambda t e^{-\lambda t}$$

The pdf of T_o is obtained by differentiating $F_{T_o}(t)$:

$$f_{T_o}(t) = \begin{cases} \lambda^2 t e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{6.4}$$

This is a gamma pdf ($\alpha = 2$ and $\beta = 1/\lambda$). This distribution for T_o can also be derived by a moment generating function argument.

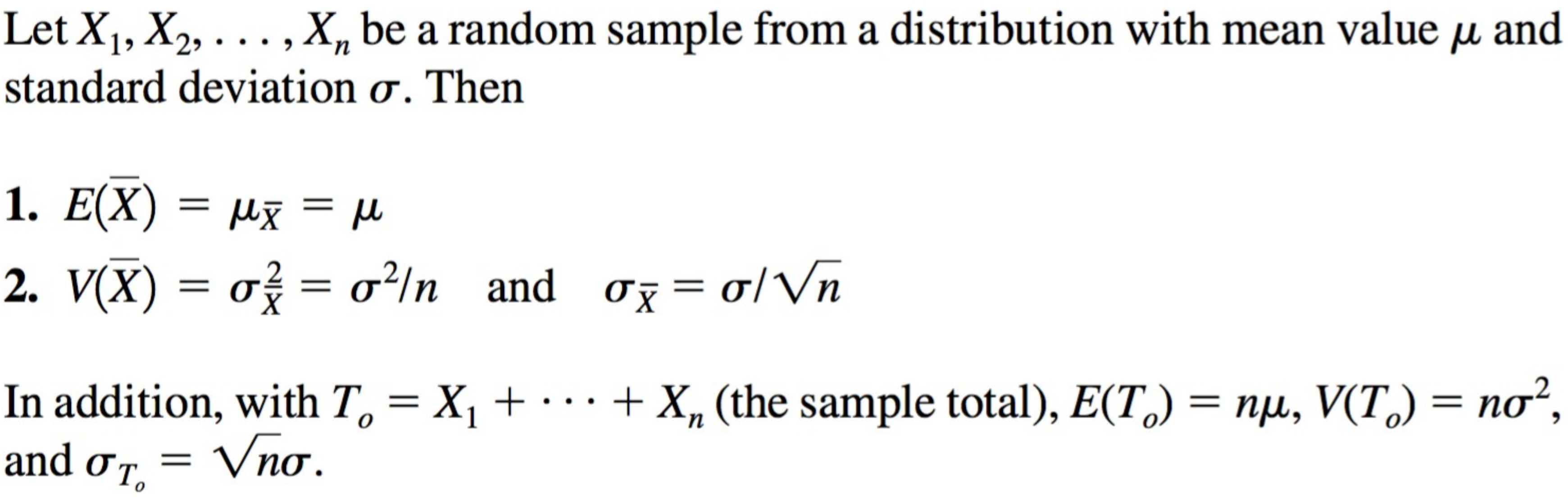
The pdf of $\overline{X} = T_o/2$ can be obtained by the method of Section 4.7 as

$$f_{\overline{X}}(\overline{x}) = \begin{cases} 4\lambda^2 \overline{x} e^{-2\lambda\overline{x}} & \overline{x} \geq 0 \\ 0 & \overline{x} < 0 \end{cases} \tag{6.5}$$

The mean and variance of the underlying exponential distribution are $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$. Using Expressions (6.4) and (6.5), it can be verified that $E(\overline{X}) = 1/\lambda$, $V(\overline{X}) = 1/(2\lambda^2)$, $E(T_o) = 2/\lambda$, and $V(T_o) = 2/\lambda^2$. These results again suggest some general relationships between means and variances of $\overline{X}$, T_o, and the underlying distribution.

The Distribution of the Sample Mean $\bar{X}$

Mean and Variance of Sample Mean



Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$
2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$

In addition, with $T_o = X_1 + \cdots + X_n$ (the sample total), $E(T_o) = n\mu$, $V(T_o) = n\sigma^2$, and $\sigma_{T_o} = \sqrt{n}\sigma$.

Proofs are deferred to the next section.

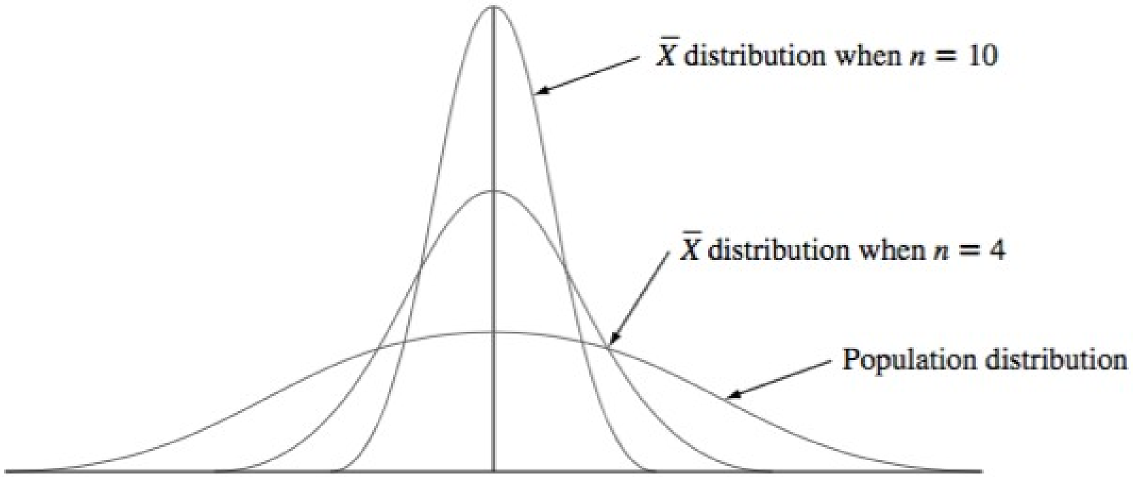
Distribution of the Mean of Normal R.V.'s

Let $X_1, \ldots, X_n$ be a random sample (IID) from $N(\mu, \sigma^2)$, then

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

$$T \sim N(n\mu, n\sigma^2)$$

where $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, T = n\bar{X}$.



Example:

The time that it takes a randomly selected rat of a certain subspecies to find its way through a maze is a normally distributed rv with $\mu = 1.5$ min and $\sigma = .35$. Suppose five rats are selected. Let $X_1, ..., X_5$ denote their times in the maze. Assuming the X_i's to be a random sample from this normal distribution, what is the probability that the mean time $\bar{X} = \frac{X_1 + ... + X_5}{5}$ for the five is smaller than 2min?

Solution:

$$\mu_{\bar{X}} = 1.5, \quad \sigma_{\bar{X}} = 0.35/\sqrt{5} = 0.1565$$

$$P(\bar{X} \leq 2.0) = P\left(Z \leq \frac{2.0 - 1.5}{.1565}\right) = P(Z \leq 3.19) = \Phi(3.19) = .9993$$

The Central Limit Theorem (CLT)

Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit as $n \to \infty$, the standardized version of $\bar{X}$ has the standard normal distribution. That is,

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \Phi(z)$$

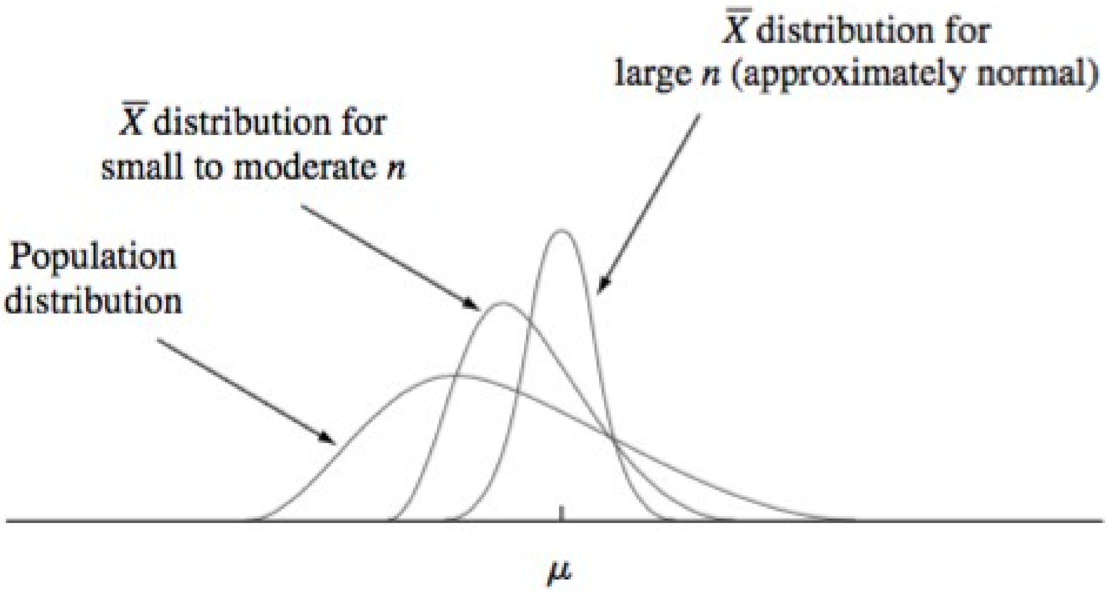
That is,

$$\bar{X} \sim (\text{approximately}) N(\mu, \sigma_{\bar{X}}^2)$$

where $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$, or $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Often we say that $\bar{X}$ is **asymptotically normal**.

Graphical demonstration of CLT


$\overline{X}$ distribution for large n (approximately normal)
$\overline{X}$ distribution for small to moderate n
Population distribution
μ

Simulation Demonstration Using R

Example:

A certain consumer organization customarily reports the number of major defects for each new automobile that it tests. Suppose the number of such defects for a certain model is a random variable with mean value 3.2 and standard deviation 2.4. Among 100 randomly selected cars of this model, how likely is it that the sample average number of major defects exceeds 4?

Solution:

Let X_i denote the number of major defects for the *i*th car in the random sample. By CLT, approximately

$$\bar{X} = \frac{X_1 + \ldots + X_{100}}{100} \sim N(3.2, 2.4/\sqrt{100})$$

$$P(\bar{X} > 4) \approx P\left(Z > \frac{4 - 3.2}{.24}\right) = 1 - \Phi(3.33) = .0004$$

A special case of CLT applied to binomial distribution

For n binomial trials with success rate p, define X_i as follows:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial results in a success} \\ 0 & \text{if the } i\text{th trial results in a failure} \end{cases} \qquad (i = 1, \ldots, n)$$

$X_i \sim \text{Bernoulli}(p)$, with mean $\mu = p,$ variance $\sigma^2 = p(1-p)$

$X = X_1 + \ldots, X_n \sim \text{Binomial}(n, p)$.

By CLT, approximately $\bar{X} = \frac{X}{n} \equiv \hat{p} \sim N(p, p(1-p)/n)$.

Example:

Maureen Webster, who is running for mayor in a large city, claims that she is favored by 53% of all eligible voters of that city. Assume that this claim is true. What is the probability that in a random sample of 400 registered voters taken from this city, less than 49% will favor Maureen Webster?

Solution:

$\hat{p} = \frac{X_1 + \ldots + X_n}{n}$, where X_i ~ Bernoulli (0.53).

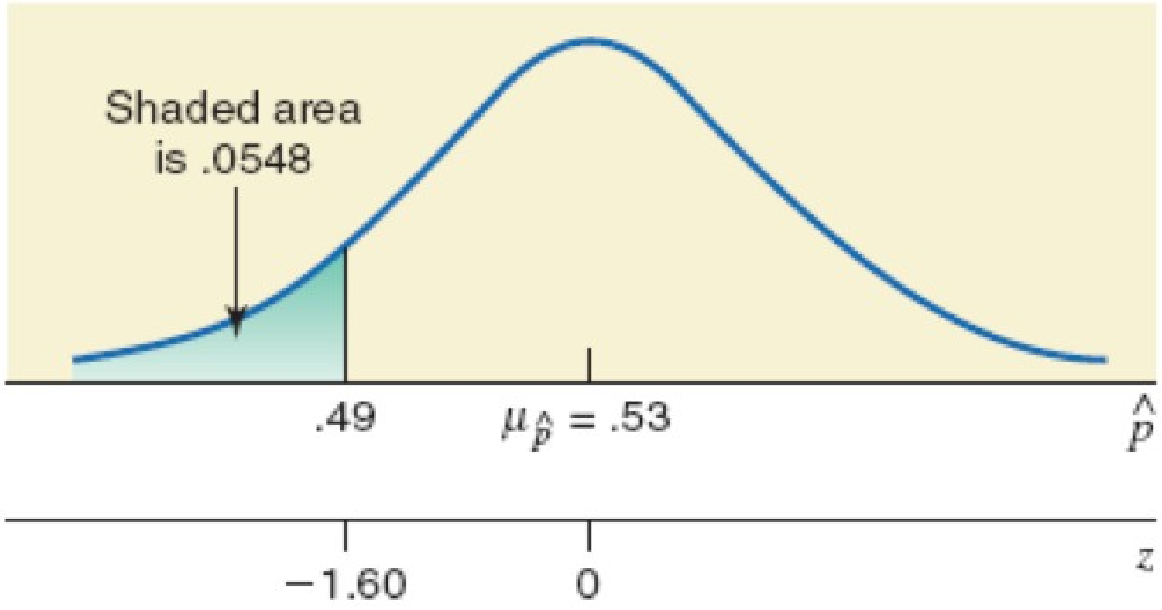
$p = 0.53$, $n = 400$.

By CLT, $\hat{p}$ is approximately distributed with mean and variance:

$$\mu_{\hat{p}} = 0.53, \sigma^2_{\hat{p}} = \frac{0.53 \times 0.47}{400} = 0.025$$

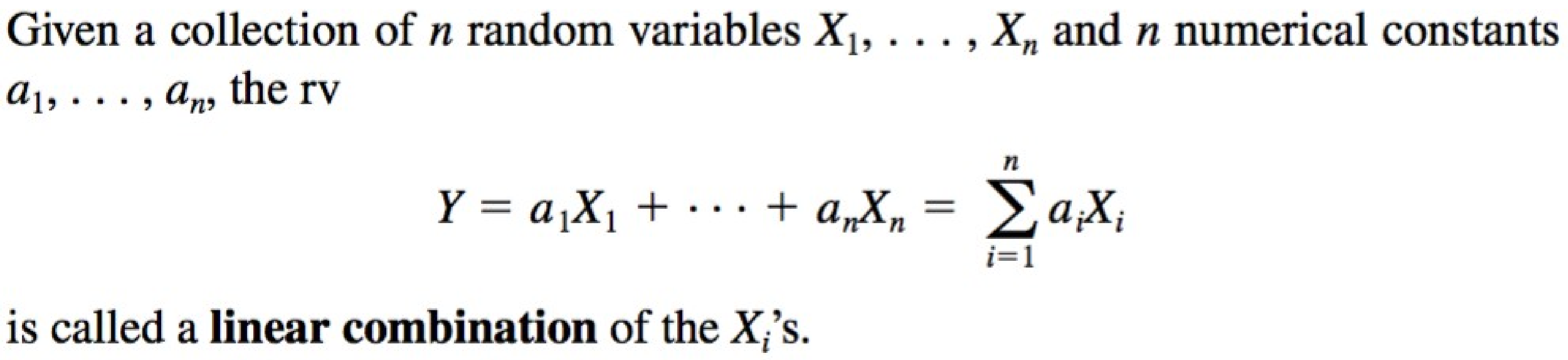
$$P(\hat{p} < .49) = P(z < -1.60)$$

$$= \mathbf{.0548}$$


Shaded area
is .0548
.49
$\mu_{\hat{p}} = .53$
$\hat{p}$
−1.60
0
z

The Distribution of a Linear Combination

Definition



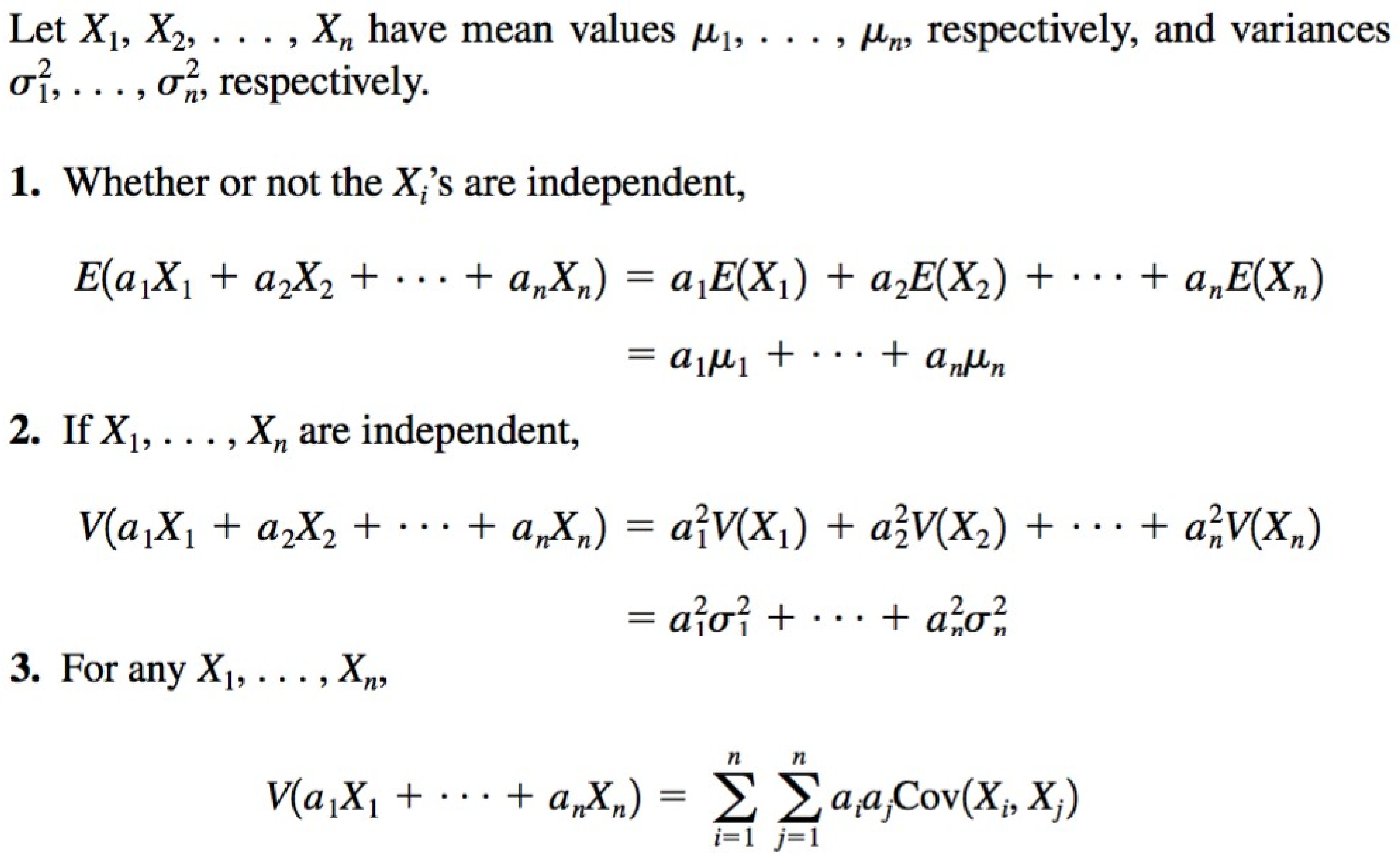
Given a collection of n random variables $X_1, \ldots, X_n$ and n numerical constants $a_1, \ldots, a_n$, the rv

$$Y = a_1X_1 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i$$

is called a **linear combination** of the X_i's.

A special case: $a_i = 1/n$, $\bar{X} = \sum_{i=1}^{n}(1/n)X_i$

Mean and Variance of a Linear Combination



Let $X_1, X_2, \ldots, X_n$ have mean values $\mu_1, \ldots, \mu_n$, respectively, and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively.

1. Whether or not the X_i's are independent,

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) \\ &= a_1\mu_1 + \cdots + a_n\mu_n \end{aligned}$$

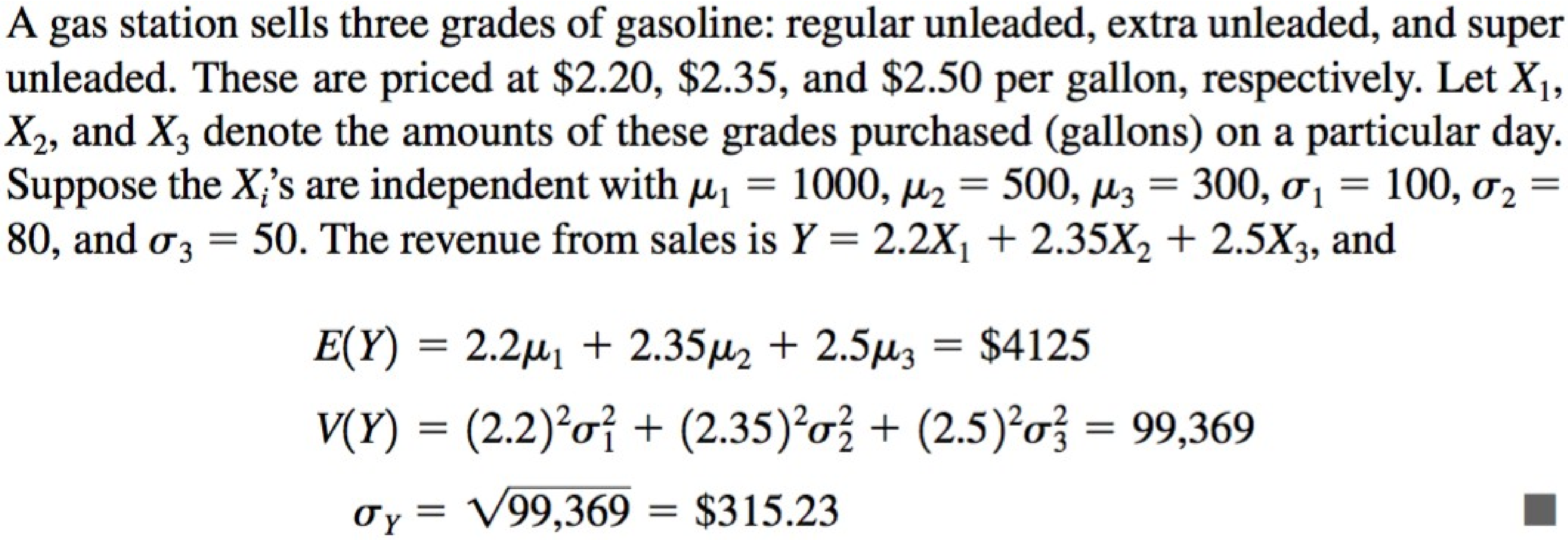
2. If $X_1, \ldots, X_n$ are independent,

$$\begin{aligned} V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) &= a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n) \\ &= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \end{aligned}$$

3. For any $X_1, \ldots, X_n$,

$$V(a_1X_1 + \cdots + a_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\text{Cov}(X_i, X_j)$$

Example



A gas station sells three grades of gasoline: regular unleaded, extra unleaded, and super unleaded. These are priced at \$2.20, \$2.35, and \$2.50 per gallon, respectively. Let X_1, X_2, and X_3 denote the amounts of these grades purchased (gallons) on a particular day. Suppose the X_i's are independent with $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300$, $\sigma_1 = 100$, $\sigma_2 = 80$, and $\sigma_3 = 50$. The revenue from sales is $Y = 2.2X_1 + 2.35X_2 + 2.5X_3$, and

$$E(Y) = 2.2\mu_1 + 2.35\mu_2 + 2.5\mu_3 = \$4125$$

$$V(Y) = (2.2)^2\sigma_1^2 + (2.35)^2\sigma_2^2 + (2.5)^2\sigma_3^2 = 99{,}369$$

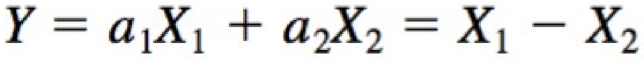
$$\sigma_Y = \sqrt{99{,}369} = \$315.23$$

■

The Difference Between Two Random Variables

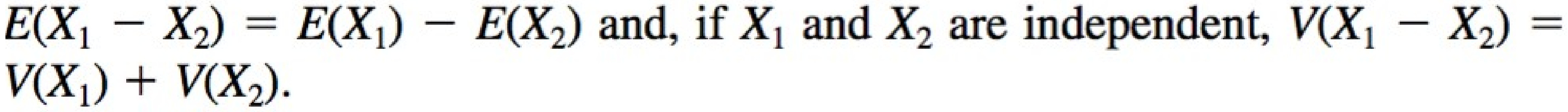
Let

$n = 2, a_1 = 1, a_2 = -1,$



$$Y = a_1X_1 + a_2X_2 = X_1 - X_2$$

We have the conclusion:



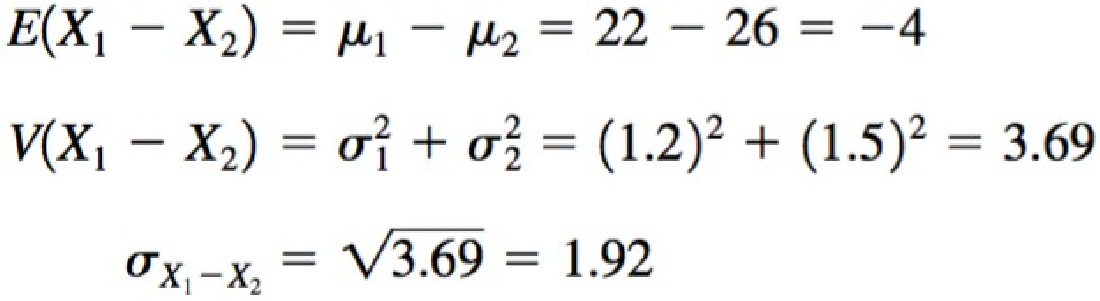
$E(X_1 - X_2) = E(X_1) - E(X_2)$ and, if X_1 and X_2 are independent, $V(X_1 - X_2) = V(X_1) + V(X_2)$.

Example

A certain automobile manufacturer equips a particular model with either a six-cylinder engine or a four-cylinder engine. Let X_1 and X_2 be fuel efficiencies for independently and randomly selected six-cylinder and four-cylinder cars, respectively. With

$$\mu_1 = 22, \mu_2 = 26, \sigma_1 = 1.2,_2 = 1.5$$

We have



$$E(X_1 - X_2) = \mu_1 - \mu_2 = 22 - 26 = -4$$

$$V(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 = (1.2)^2 + (1.5)^2 = 3.69$$

$$\sigma_{X_1 - X_2} = \sqrt{3.69} = 1.92$$

Distribution of Linear Combination of Normal R.V.'s

Theorem: Suppose X_i is normally distributed with mean μ_i and standard deviation σ_i. Assume that $X_1, \ldots, X_n$ are independent. Let

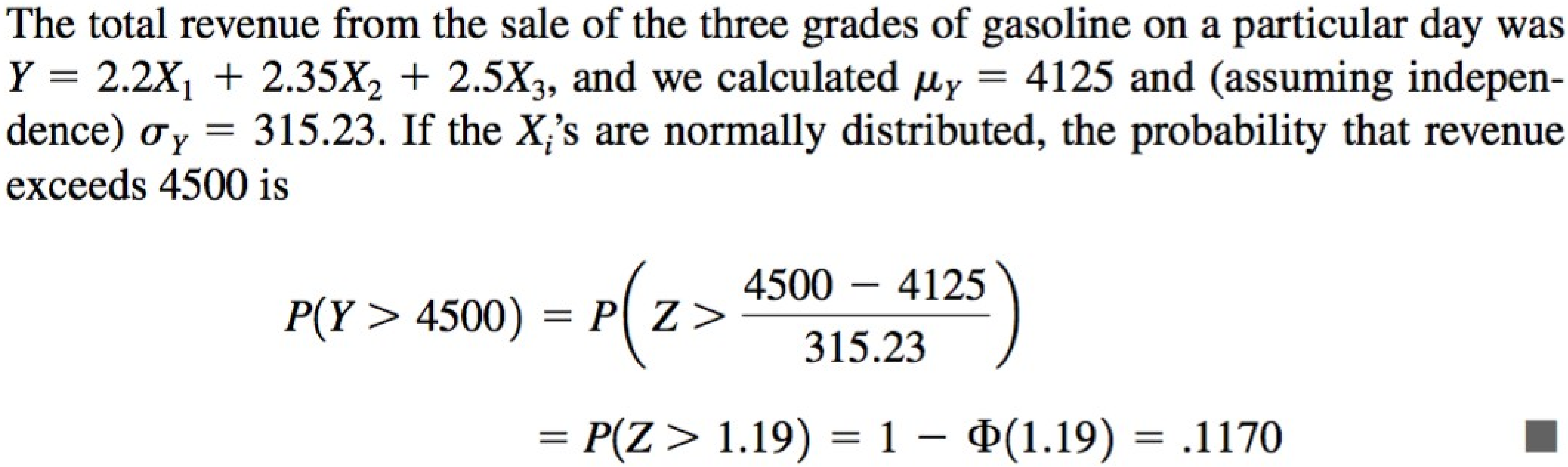
$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n.$$

Y has a normal distribution with mean and variances as follows:

$$E(Y) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

$$V(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

Example



The total revenue from the sale of the three grades of gasoline on a particular day was $Y = 2.2X_1 + 2.35X_2 + 2.5X_3$, and we calculated $\mu_Y = 4125$ and (assuming independence) $\sigma_Y = 315.23$. If the X_i's are normally distributed, the probability that revenue exceeds 4500 is

$$P(Y > 4500) = P\left(Z > \frac{4500 - 4125}{315.23}\right)$$

$$= P(Z > 1.19) = 1 - \Phi(1.19) = .1170 \qquad ■$$

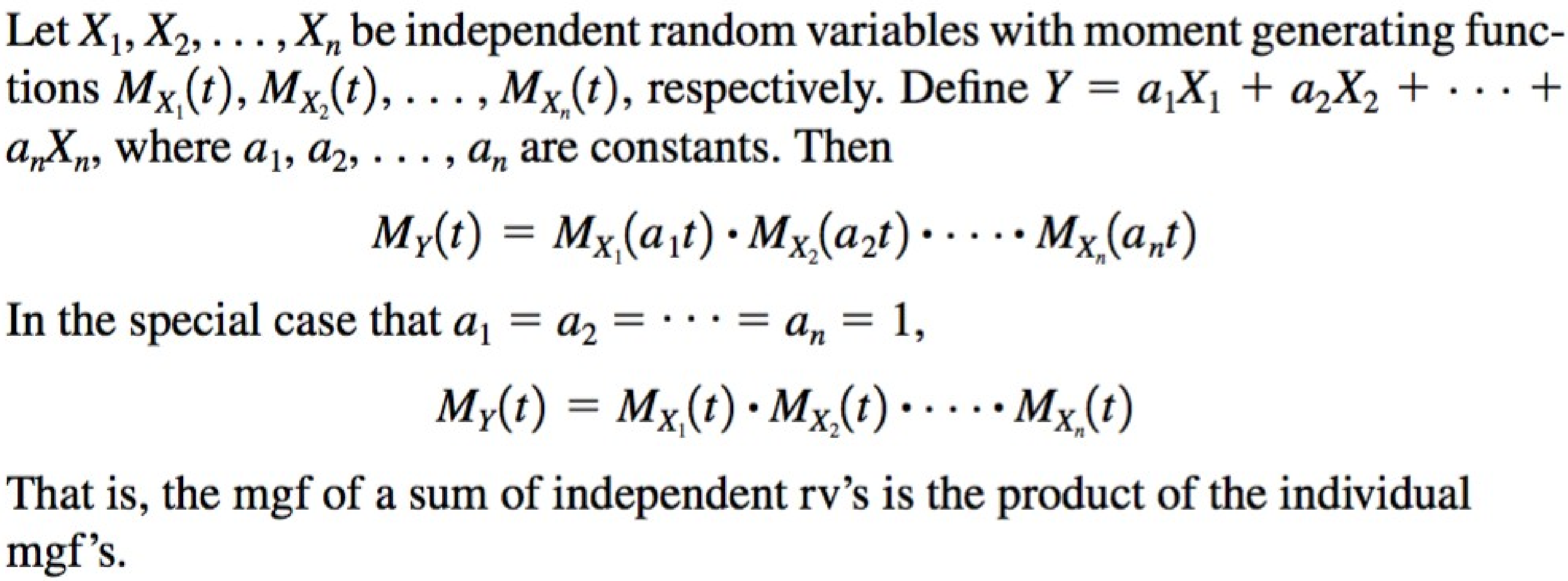
Proofs of the mean and variance of linear combination

For simplicity, $n = 2$,

$$
\begin{aligned}
E(a_1X_1 + a_2X_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(a_1x_1 + a_2x_2)f(x_1, x_2)\,dx_1\,dx_2 \\
&= a_1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1f(x_1, x_2)\,dx_2\,dx_1 + a_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_2f(x_1, x_2)\,dx_1\,dx_2 \\
&= a_1\int_{-\infty}^{\infty}x_1f_{X_1}(x_1)\,dx_1 + a_2\int_{-\infty}^{\infty}x_2f_{X_2}(x_2)\,dx_2 \\
&= a_1E(X_1) + a_2E(X_2)
\end{aligned}
$$

$$
\begin{aligned}
V(a_1X_1 + a_2X_2) &= E\{[a_1X_1 + a_2X_2 - (a_1\mu_1 + a_2\mu_2)]^2\} \\
&= E\{a_1^2(X_1 - \mu_1)^2 + a_2^2(X_2 - \mu_2)^2 + 2a_1a_2(X_1 - \mu_1)(X_2 - \mu_2)\} \\
&= a_1^2V(X_1) + a_2^2V(X_2) + 2a_1a_2\mathrm{Cov}(X_1, X_2)
\end{aligned}
$$

Moment Generating Function of a Linear Combination



Let $X_1, X_2, \dots, X_n$ be independent random variables with moment generating functions $M_{X_1}(t), M_{X_2}(t), \dots, M_{X_n}(t)$, respectively. Define $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$, where $a_1, a_2, \dots, a_n$ are constants. Then

$$M_Y(t) = M_{X_1}(a_1t) \cdot M_{X_2}(a_2t) \cdot \cdots \cdot M_{X_n}(a_nt)$$

In the special case that $a_1 = a_2 = \cdots = a_n = 1$,

$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \cdots \cdot M_{X_n}(t)$$

That is, the mgf of a sum of independent rv's is the product of the individual mgf's.

Proof:

Using definition of MGF,

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(e^{t(a_1X_1+a_2X_2+\cdots+a_nX_n)}) \\ &= E(e^{ta_1X_1+ta_2X_2+\cdots+ta_nX_n}) = E(e^{ta_1X_1} \bullet e^{ta_2X_2} \bullet \cdots \bullet e^{ta_nX_n}) \end{aligned}$$

Further,

$$\begin{aligned} E(e^{ta_1X_1} \bullet e^{ta_2X_2} \bullet \cdots \bullet e^{ta_nX_n}) &= E(e^{ta_1X_1}) \cdot E(e^{ta_2X_2}) \cdot \cdots \cdot E(e^{ta_nX_n}) \\ &= M_{X_1}(a_1t) \cdot M_{X_2}(a_2t) \cdot \cdots \cdot M_{X_n}(a_nt) \end{aligned}$$

Derivation of Distribution of A Linear Combination of Normal R.V.'s

Suppose X_i is normally distributed with mean μ_i and standard deviation σ_i. Assume that $X_1, \ldots, X_n$ are independent. Let

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Because $X_i \sim N(\mu_i, \sigma_i^2)$, the MGF of X_i is

$$M_{X_i}(t) = e^{\mu_i t + \sigma_i^2 t^2/2}$$

We can then find MGF of Y:

$$\begin{aligned} M_Y(t) &= M_{X_1}(a_1 t) \cdot M_{X_2}(a_2 t) \cdot \cdots \cdot M_{X_n}(a_n t) \\ &= e^{\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2} e^{\mu_2 a_2 t + \sigma_2^2 a_2^2 t^2/2} \cdot \cdots \cdot e^{\mu_n a_n t + \sigma_n^2 a_n^2 t^2/2} \\ &= e^{(\mu_1 a_1 + \mu_2 a_2 + \cdots + \mu_n a_n)t + (\sigma_1^2 a_1^2 + \sigma_2^2 a_2^2 + \cdots + \sigma_n^2 a_n^2)t^2/2} \end{aligned}$$

Because the moment generating function of *Y* is the moment generating function of a normal random variable, it follows that *Y* is normally distributed by the uniqueness principle for MGF. The mean and variance are given as follows:

$$E(Y) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$$

$$V(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$$

Distributions Based on a Normal Random Sample

The Chi-Squared Distribution

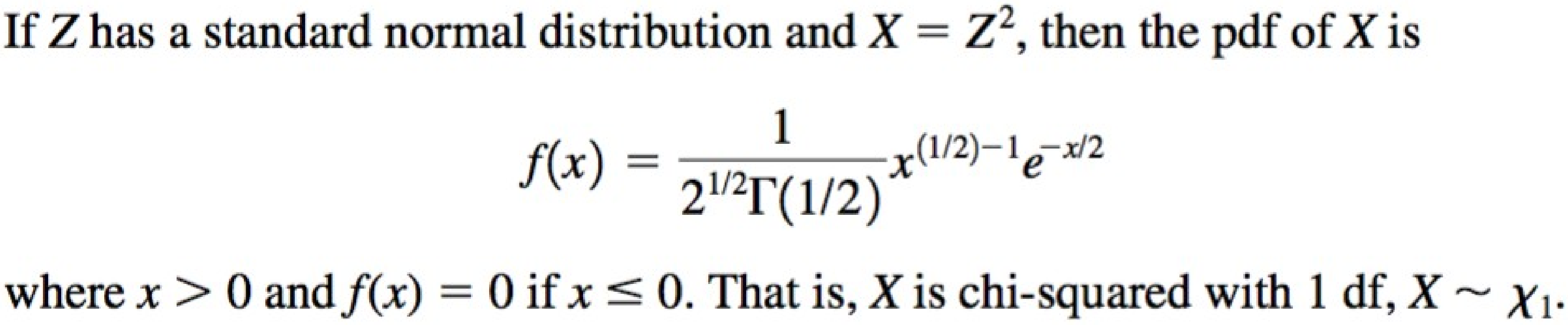
The chi-squared distribution is a special case of the gamma distribution. It has one parameter n called the number of degrees of freedom of the distribution. Possible values of ν are 1, 2, 3, ... The chi-squared pdf is

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1}\, e^{-x/2} \quad \text{if } x > 0,\ f(x) = 0 \quad \text{if } x \leq 0$$

Mean, variance and MGF:

$$\mu = \alpha\beta = \nu \qquad \sigma^2 = \alpha\beta^2 = 2\nu \qquad M_X(t) = (1 - 2t)^{-\nu/2}$$

Distribution of Squared Normal

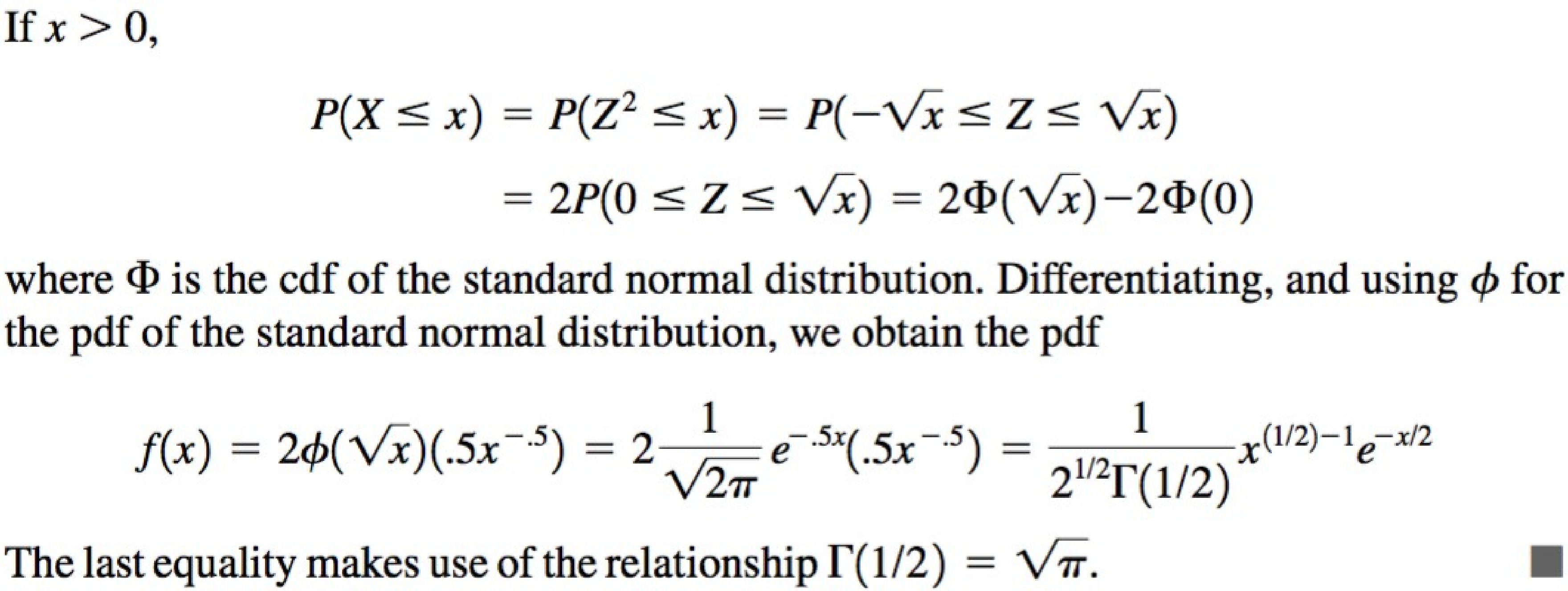


If Z has a standard normal distribution and $X = Z^2$, then the pdf of X is

$$f(x) = \frac{1}{2^{1/2}\Gamma(1/2)} x^{(1/2)-1} e^{-x/2}$$

where $x > 0$ and $f(x) = 0$ if $x \leq 0$. That is, X is chi-squared with 1 df, $X \sim \chi_1$.

Proof:



If $x > 0$,

$$\begin{aligned} P(X \leq x) &= P(Z^2 \leq x) = P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= 2P(0 \leq Z \leq \sqrt{x}) = 2\Phi(\sqrt{x}) - 2\Phi(0) \end{aligned}$$

where Φ is the cdf of the standard normal distribution. Differentiating, and using ϕ for the pdf of the standard normal distribution, we obtain the pdf

$$f(x) = 2\phi(\sqrt{x})(.5x^{-.5}) = 2\frac{1}{\sqrt{2\pi}} e^{-.5x}(.5x^{-.5}) = \frac{1}{2^{1/2}\Gamma(1/2)} x^{(1/2)-1} e^{-x/2}$$

The last equality makes use of the relationship $\Gamma(1/2) = \sqrt{\pi}$. ■

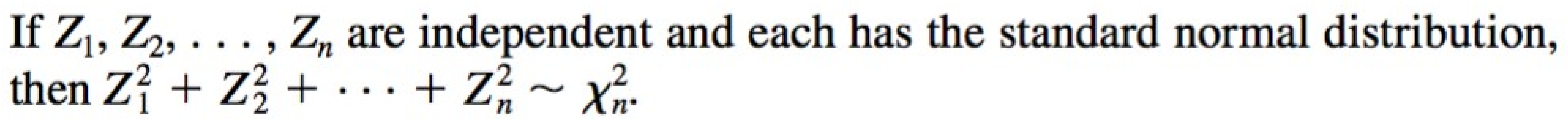
Additivity of Chi-Squared Distribution

If $X_1 \sim \chi^2_{\nu_1}$, $X_2 \sim \chi^2_{\nu_2}$, and they are independent, then $X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2}$.

Proof: If random variables are independent, then the moment generating function of their sum is the product of their moment generating functions. Therefore,

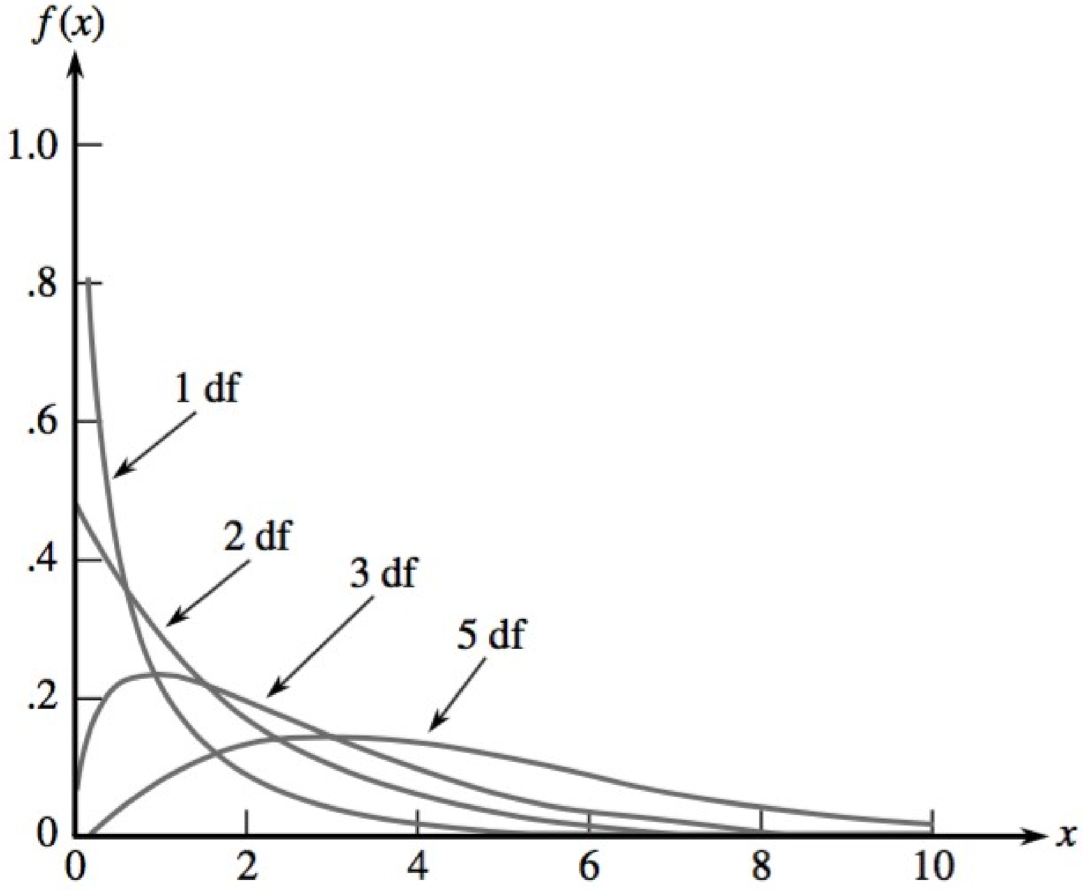
$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = (1 - 2t)^{-\nu_1/2}(1 - 2t)^{-\nu_2/2} = (1 - 2t)^{-(\nu_1+\nu_2)/2}$$

Sum of Squared Normal R.V.'s



If $Z_1, Z_2, \ldots, Z_n$ are independent and each has the standard normal distribution, then $Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi_n^2$.

Note that $\chi_n^2 = \mathrm{Gamma}(\alpha = n/2, \beta = 2)$.



Chi-squared density curves

Upper quantiles of Chi-Squared Distribution

$$P(\chi_2^2 > 9.210) = .01 \qquad \chi_{.01,2}^2 = 9.210$$

$$P(\chi_\nu^2 > c) = \alpha \qquad \chi_{\alpha,\nu}^2 = c$$

Learn to check chi-squared table.

Student's Theorem

Suppose $X_1, X_2, ..., X_n$ are a random sample from $N(\mu, \sigma^2)$. Let

$$\bar{X} = \frac{X_1 + ... + X_n}{n},$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

We have:

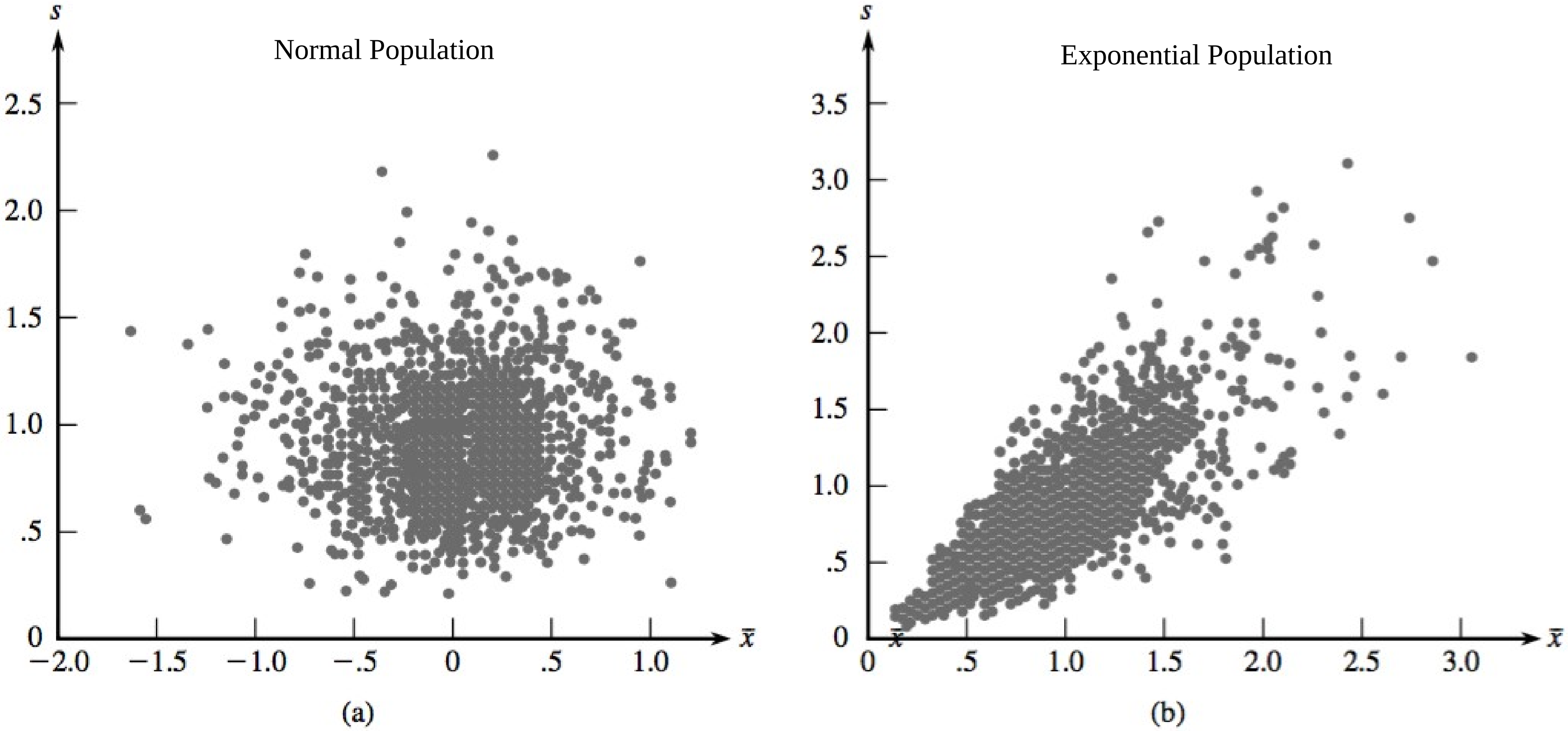
1) $\bar{X} \sim N(\mu, \sigma^2/n)$

2) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

3) $\bar{X}$ and S^2 are independent.

Proof: on blackboard.

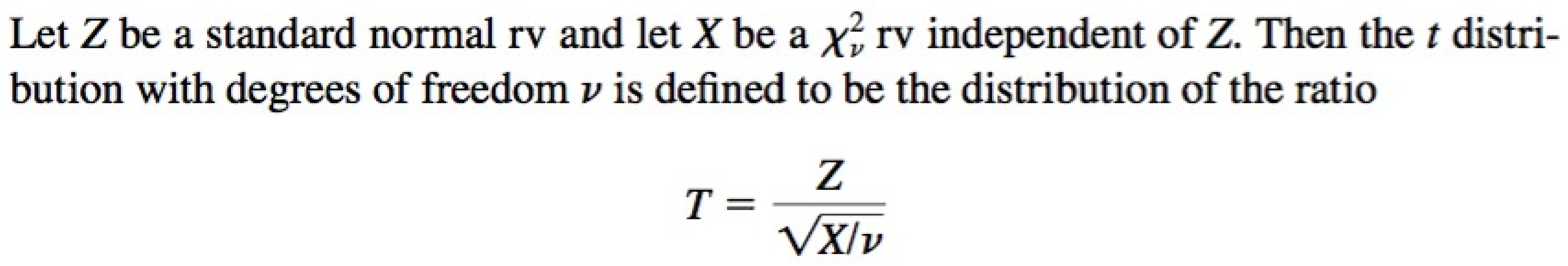
A Simulation Demonstration of Student's Theorem



Scatter plot of ($\bar{x}$, s) pairs

The t Distribution

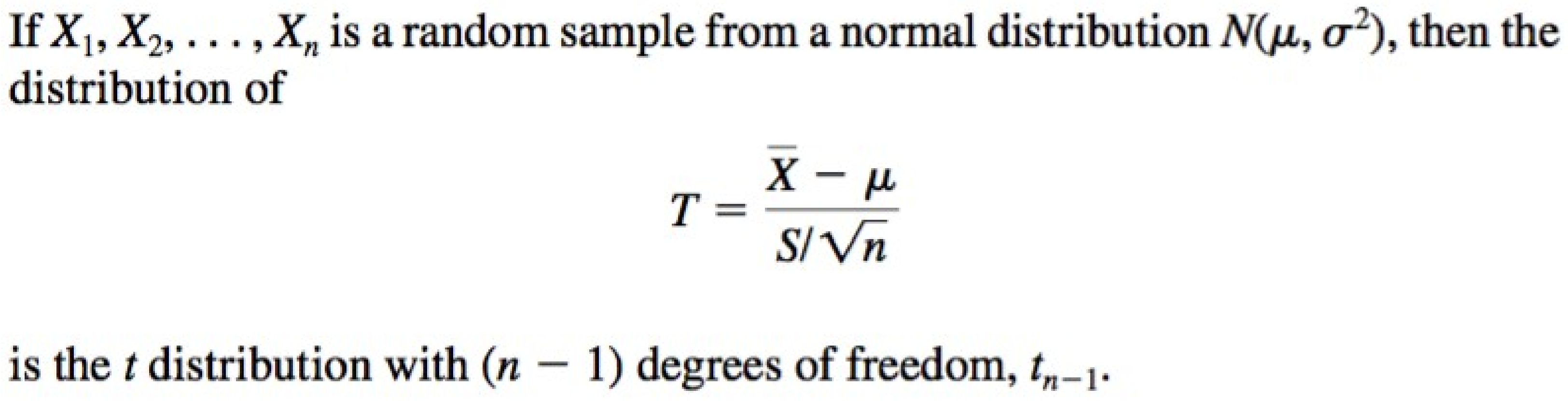
Definition:



Let Z be a standard normal rv and let X be a χ^2_ν rv independent of Z. Then the t distribution with degrees of freedom ν is defined to be the distribution of the ratio

$$T = \frac{Z}{\sqrt{X/\nu}}$$

A Theorem about Normal Sample

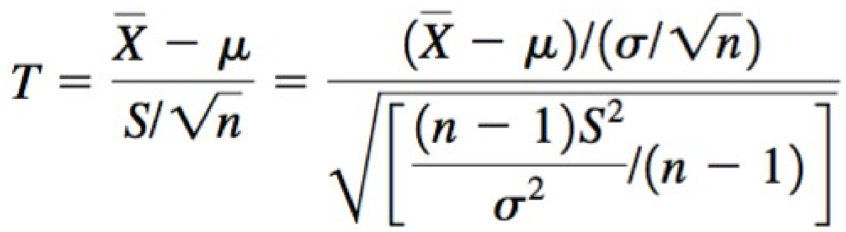


If $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, then the distribution of

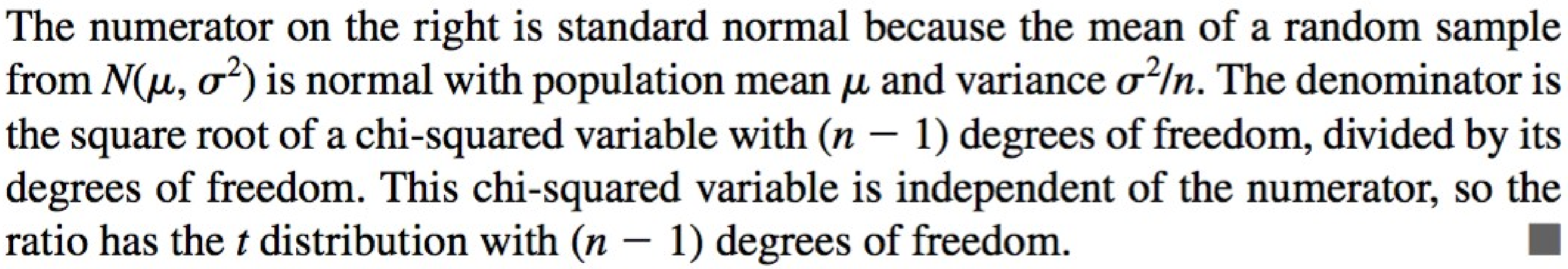
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is the t distribution with $(n - 1)$ degrees of freedom, t_{n-1}.

Proof:

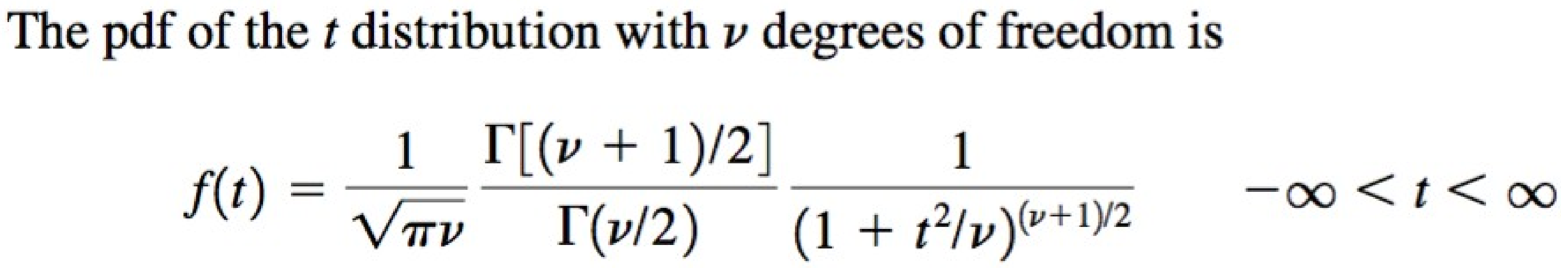


$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\left[\frac{(n-1)S^2}{\sigma^2}/(n-1)\right]}}$$



The numerator on the right is standard normal because the mean of a random sample from $N(\mu, \sigma^2)$ is normal with population mean μ and variance σ^2/n. The denominator is the square root of a chi-squared variable with $(n - 1)$ degrees of freedom, divided by its degrees of freedom. This chi-squared variable is independent of the numerator, so the ratio has the t distribution with $(n - 1)$ degrees of freedom. ■

PDF of t_ν

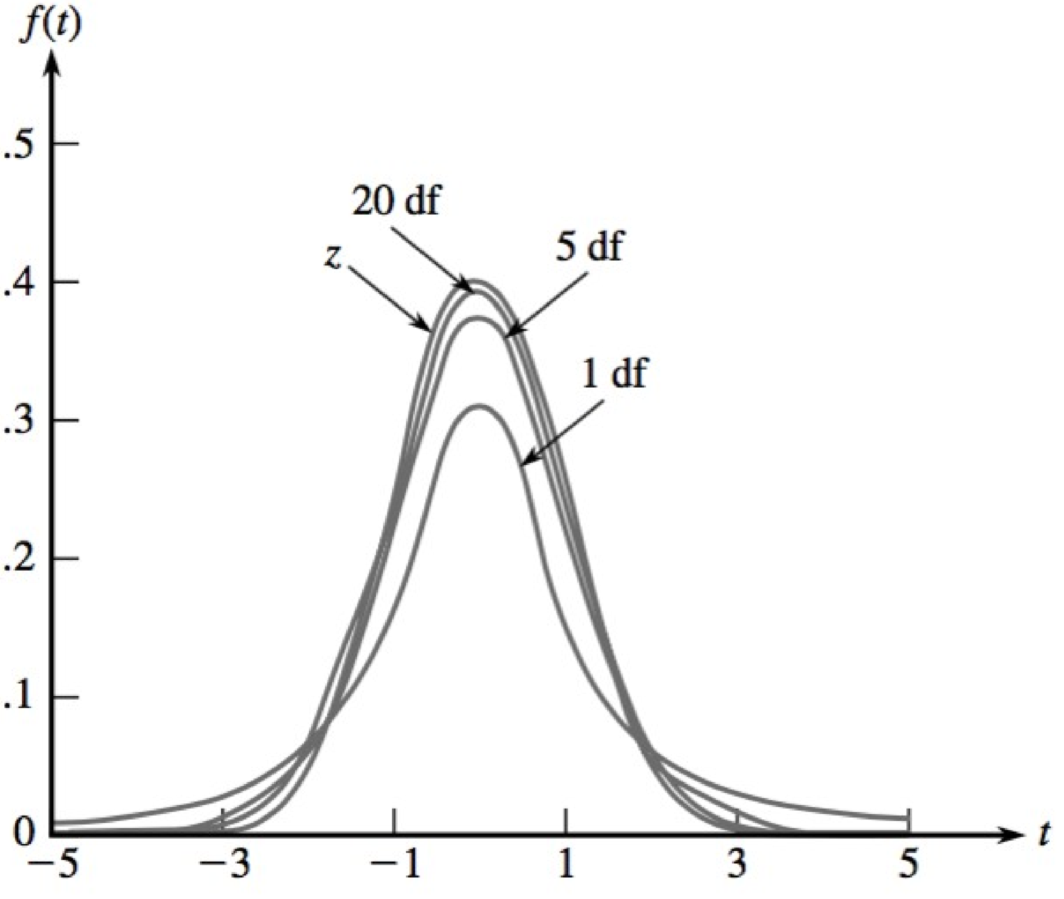


The pdf of the t distribution with ν degrees of freedom is

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma[(\nu + 1)/2]}{\Gamma(\nu/2)} \frac{1}{(1 + t^2/\nu)^{(\nu+1)/2}} \qquad -\infty < t < \infty$$

Proof: Using transformed random variables formulae.

Shape of PDF of t_ν

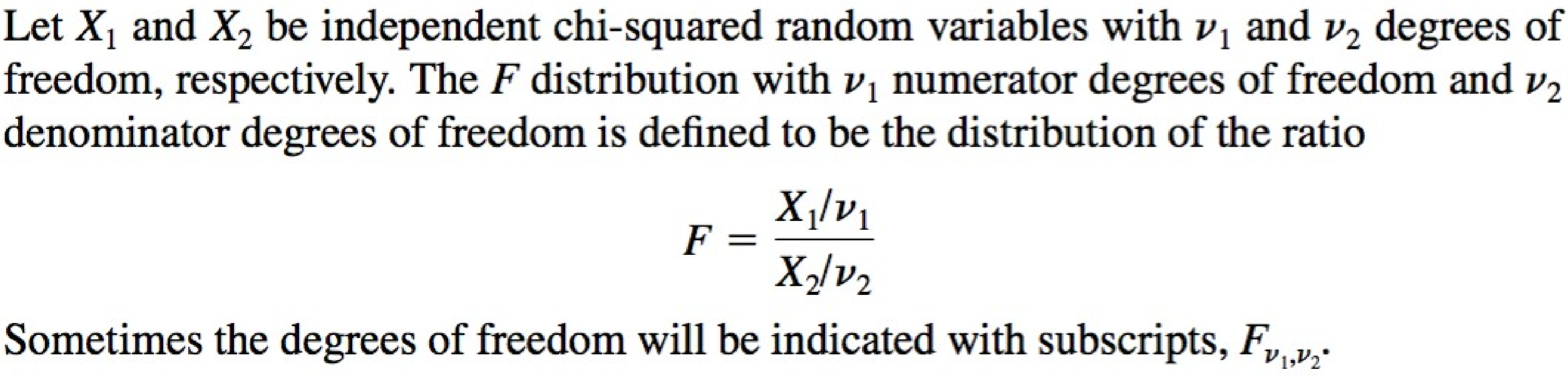


Comparison of *t* curves to the *z* curve

$E(T) = 0$ if $\nu > 1$

$V(T) = \frac{\nu}{\nu-2}$ if $\nu > 2$

The F Distribution

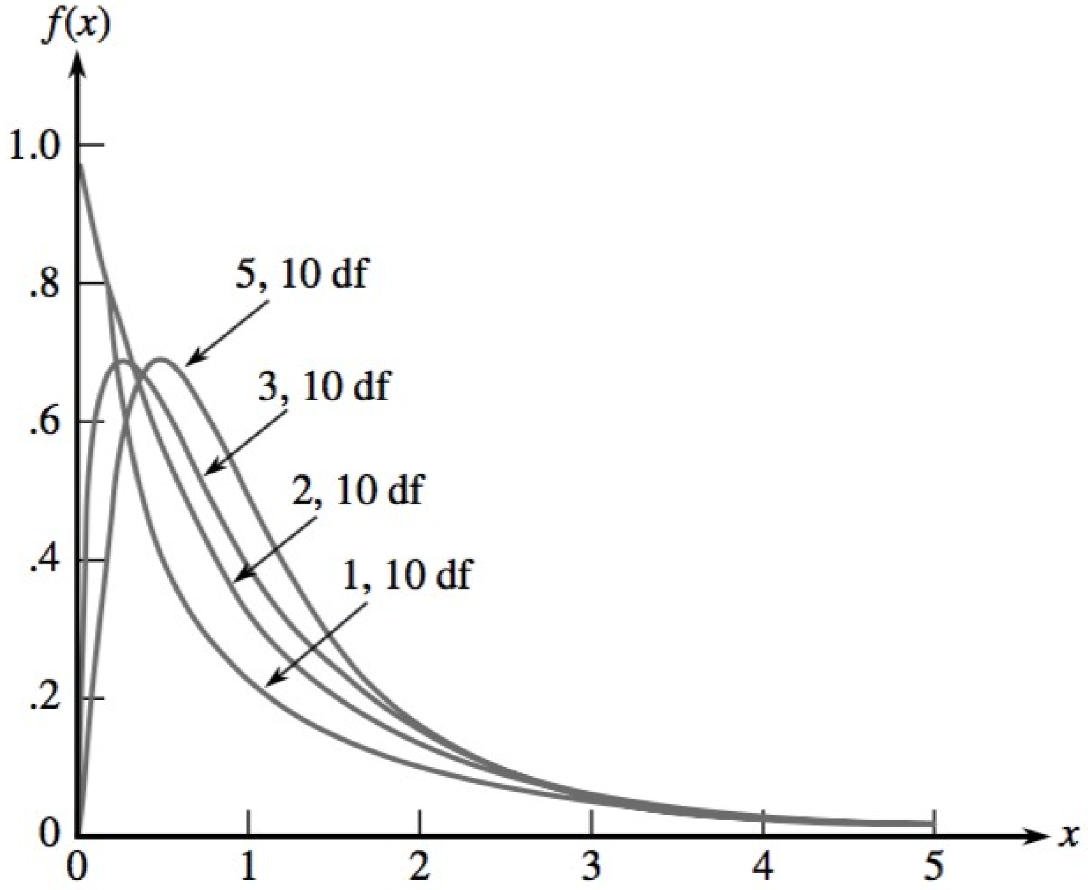


Let X_1 and X_2 be independent chi-squared random variables with ν_1 and ν_2 degrees of freedom, respectively. The F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom is defined to be the distribution of the ratio

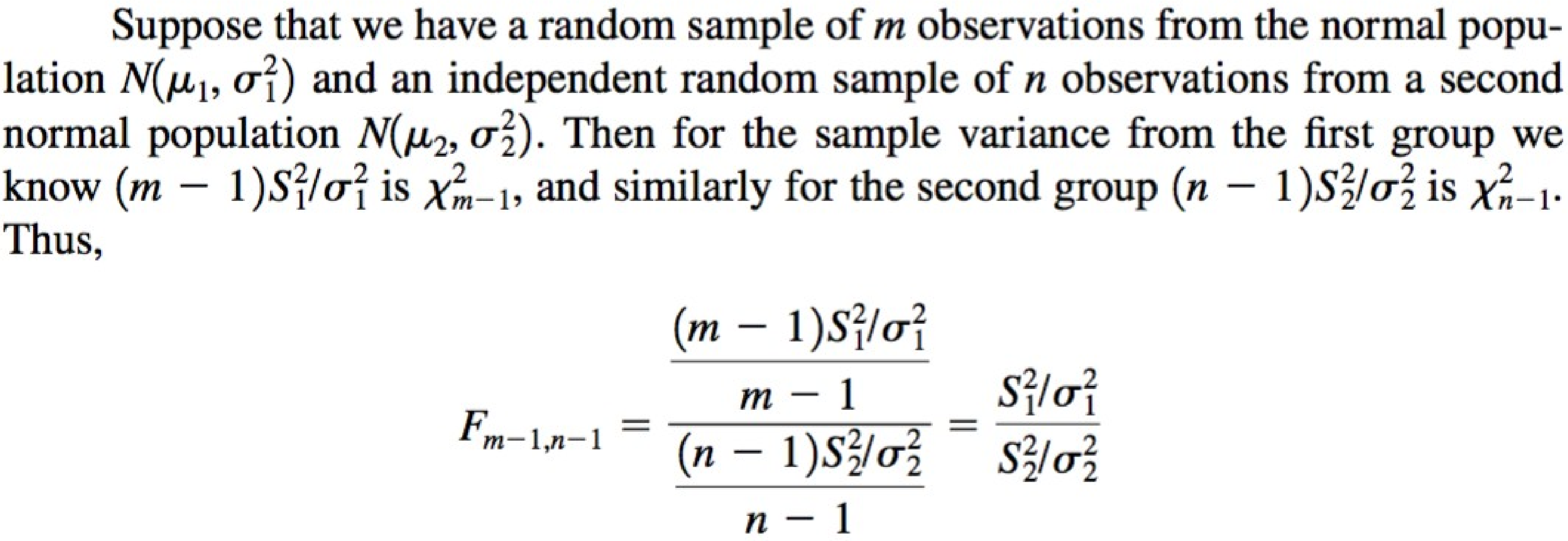
$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

Sometimes the degrees of freedom will be indicated with subscripts, F_{ν_1,ν_2}.

PDF Curves of F Distributions



An Application to Normal Samples



Suppose that we have a random sample of m observations from the normal population $N(\mu_1, \sigma_1^2)$ and an independent random sample of n observations from a second normal population $N(\mu_2, \sigma_2^2)$. Then for the sample variance from the first group we know $(m-1)S_1^2/\sigma_1^2$ is χ^2_{m-1}, and similarly for the second group $(n-1)S_2^2/\sigma_2^2$ is χ^2_{n-1}. Thus,

$$F_{m-1,n-1} = \frac{\dfrac{(m-1)S_1^2/\sigma_1^2}{m-1}}{\dfrac{(n-1)S_2^2/\sigma_2^2}{n-1}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has F distribution with degree freedoms *m*-1 and *n*-1.

Upper Quantiles of F Distribution

F_{α,ν_1,ν_2} Is the value c such that $P(F_{\nu_1,\nu_2} > c) = \alpha$.