Applied Statistics

Fundamental Sampling Distributions

Jiebo Song, Assistant Professor

Beijing Institute
of Mathematical
Sciences and Applications

Oct. 8, 2024

Outline

Some Discrete Probability Distributions

- Binomial and Multinomial Distributions
- Hypergeometric Distribution
- Negative Binomial and Geometric Distributions
- Poisson distribution and the Poisson Process

Some Continuous Probability Distributions

- Continuous Uniform Distribution / Normal Distribution
- Gamma and Exponential Distributions
- Chi-Squared Distribution / Beta Distribution / Lognormal Distribution / Weibull Distribution

Fundamental Sampling Distributions and Data Descriptions

- Some Important Statistics / Sampling Distributions
- t-Distribution / F-Distribution

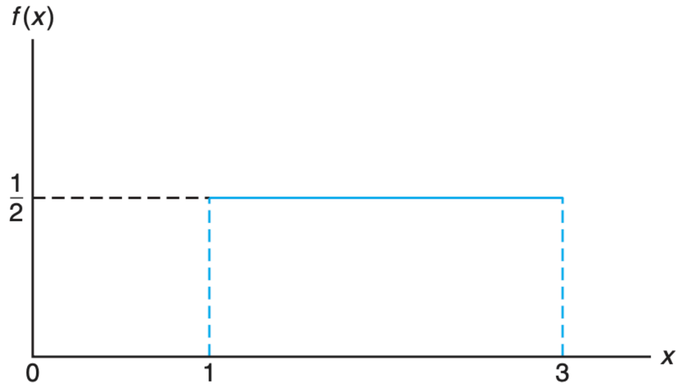
Some Continuous Probability Distributions

Continuous Uniform Distribution

- Rectangular distribution

The density function of the continuous uniform random variable X on the interval $[A, B]$ is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \le x \le B, \\ 0, & \text{elsewhere.} \end{cases}$$



The mean and variance of the uniform distribution are

$$\mu = \frac{A+B}{2} \text{ and } \sigma^2 = \frac{(B-A)^2}{12}.$$

Some Continuous Probability Distributions

Normal Distribution

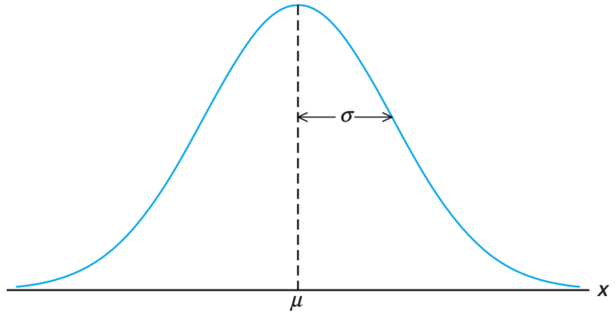
https://web.stanford.edu/class/archive/cs/cs109/cs109.1226/

- Normal curve / Gaussian distribution

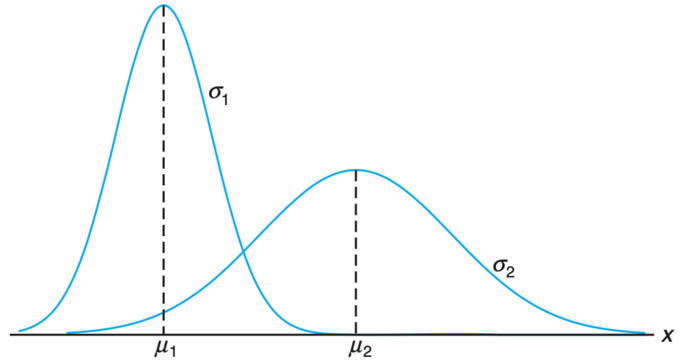
The density of the normal random variable X, with mean μ and variance σ^2, is

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty,$$

where $\pi = 3.14159\ldots$ and $e = 2.71828\ldots$.



The mean and variance of $n(x;\mu,\sigma)$ are μ and σ^2, respectively. Hence, the standard deviation is σ.



- **The curve has its points of inflection at x = μ ± σ; it is concave downward if μ − σ < X < μ + σ and is concave upward otherwise.**
- **Under certain conditions, the normal distribution provides a good continuous approximation to the binomial and hypergeometric distributions.**
- **The limiting distribution of sample averages is normal.**

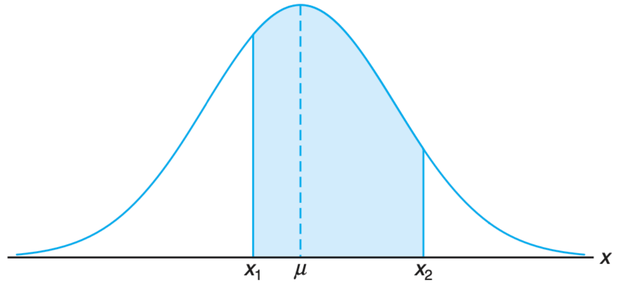
Some Continuous Probability Distributions

Normal Distribution

https://web.stanford.edu/class/archive/cs/cs109/cs109.1226/

- Normal curve / Gaussian distribution

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x;\mu,\sigma)\,dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx$$



The distribution of a normal random variable with mean 0 and variance 1 is called a **standard normal distribution**.

- **The difficulty encountered in solving integrals of normal density functions necessitates the tabulation of normal curve areas for quick reference.**
- **However, it would be a hopeless task to attempt to set up separate tables for every conceivable value of μ and σ.**
- **Fortunately, we are able to transform all the observations of any normal random variable X into a new set of observations of a normal random variable Z with mean 0 and variance 1.**

$$Z = \frac{X-\mu}{\sigma}. \quad z_1 = (x_1-\mu)/\sigma \quad z_2 = (x_2-\mu)/\sigma$$

$$P(x_1 < X < x_2) = \frac{1}{\sqrt{2\pi}\sigma}\int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx = \frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}\,dz$$
$$= \int_{z_1}^{z_2} n(z;0,1)\,dz = P(z_1 < Z < z_2),$$

Some Continuous Probability Distributions

Normal Approximation to the Binomial

If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of

$$Z = \frac{X - np}{\sqrt{npq}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

- **It turns out that the normal distribution with μ = np and σ² = np(1 – p) not only provides a very accurate approximation to the binomial distribution when n is large and p is not extremely close to 0 or 1 but also provides a fairly good approximation even when n is small and p is reasonably close to 1/2.**
- **The normal approximation is most useful in calculating binomial sums for large values of n.**

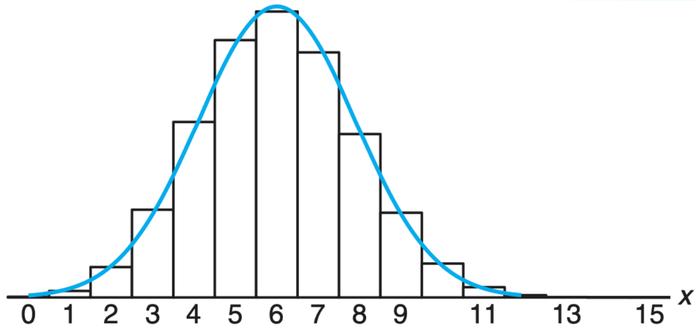


Figure 6.22: Normal approximation of $b(x; 15, 0.4)$.

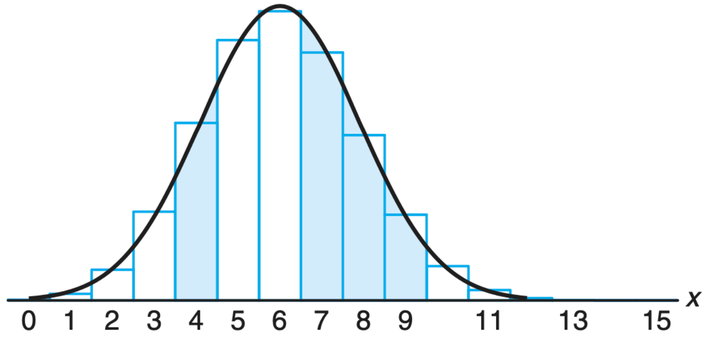


Figure 6.23: Normal approximation of $b(x; 15, 0.4)$ and $\sum_{x=7}^{9} b(x; 15, 0.4)$.

$$P(7 \le X \le 9) = \sum_{x=0}^{9} b(x; 15, 0.4) - \sum_{x=0}^{6} b(x; 15, 0.4)$$
$$= 0.9662 - 0.6098 = 0.3564,$$

$$z_1 = \frac{6.5 - 6}{1.897} = 0.26 \quad \text{and} \quad z_2 = \frac{9.5 - 6}{1.897} = 1.85.$$

$$P(7 \le X \le 9) \approx P(0.26 < Z < 1.85) = P(Z < 1.85) - P(Z < 0.26)$$
$$= 0.9678 - 0.6026 = 0.3652.$$

Some Continuous Probability Distributions

Normal Approximation to the Binomial

If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of

$$Z = \frac{X - np}{\sqrt{npq}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Let X be a binomial random variable with parameters n and p. For large n, X has approximately a normal distribution with $\mu = np$ and $\sigma^2 = npq = np(1-p)$ and

$$\begin{aligned} P(X \le x) &= \sum_{k=0}^{x} b(k; n, p) \\ &\approx \text{area under normal curve to the left of } x + 0.5 \\ &= P\left(Z \le \frac{x + 0.5 - np}{\sqrt{npq}}\right), \end{aligned}$$

and the approximation will be good if np and $n(1-p)$ are greater than or equal to 5.

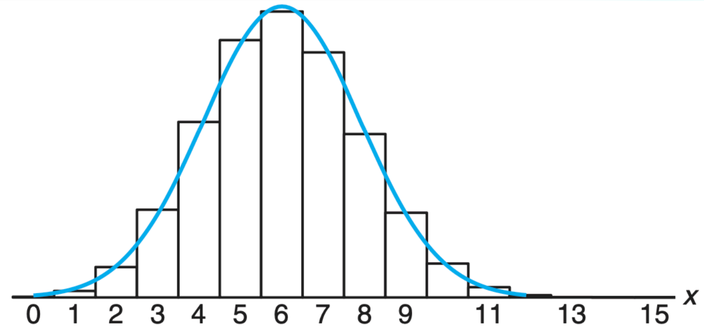


Figure 6.22: Normal approximation of $b(x; 15, 0.4)$.

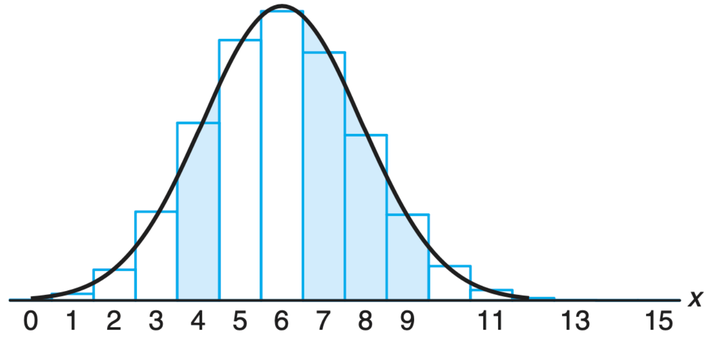


Figure 6.23: Normal approximation of $b(x; 15, 0.4)$ and $\sum_{x=7}^{9} b(x; 15, 0.4)$.

Some Continuous Probability Distributions

Gamma and Exponential Distribution

- The gamma function:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \qquad \text{for } \alpha > 0.$$

(a) $\Gamma(n) = (n-1)(n-2)\cdots(1)\Gamma(1)$, for a positive integer n.

(b) $\Gamma(n) = (n-1)!$ for a positive integer n.

(c) $\Gamma(1) = 1$.

(d) $\Gamma(1/2) = \sqrt{\pi}$.

- **The exponential and gamma distributions play an important role in both queuing theory and reliability problems.**
- **Time between arrivals at service facilities and time to failure of component parts and electrical systems often are nicely modeled by the exponential distribution.**

Some Continuous Probability Distributions

Gamma and Exponential Distribution

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \qquad \text{for } \alpha > 0.$$

- The gamma distribution:

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & x > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$.

The mean and variance of the gamma distribution are

$$\mu = \alpha\beta \text{ and } \sigma^2 = \alpha\beta^2.$$

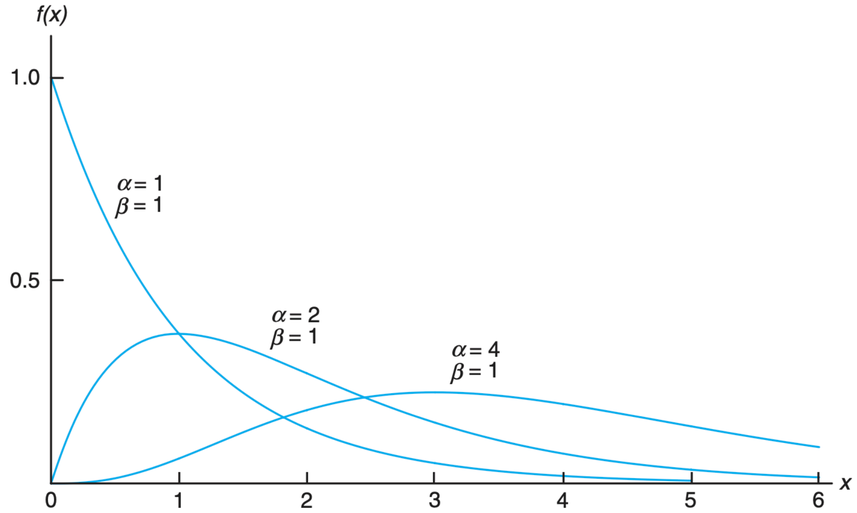


Figure 6.28: Gamma distributions.

- The exponential distribution: $\alpha = 1$

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

The mean and variance of the exponential distribution are

$$\mu = \beta \text{ and } \sigma^2 = \beta^2.$$

- **The relationship between the exponential distribution (often called the negative exponential) and the Poisson process is quite simple.**

Some Continuous Probability Distributions

Chi-Squared Distribution

- Another very important special case of the gamma distribution is obtained by letting α = v/2 and β = 2, where v is a positive integer.

The continuous random variable X has a **chi-squared distribution**, with v **degrees of freedom**, if its density function is given by

$$f(x;v) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}, & x > 0, \\ 0, & \text{elsewhere}, \end{cases}$$

where v is a positive integer.

The mean and variance of the chi-squared distribution are

$$\mu = v \text{ and } \sigma^2 = 2v.$$

- **The chi-squared distribution plays a vital role in statistical inference.**
- **It has considerable applications in both methodology and theory.**
- **The chi-squared distribution is an important component of statistical hypothesis testing and estimation.**

Some Continuous Probability Distributions

Beta Distribution

- An extension to the uniform distribution is a beta distribution.

The continuous random variable X has a **beta distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its density function is given by

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

A **beta function** is defined by

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \text{ for } \alpha,\beta > 0,$$

where $\Gamma(\alpha)$ is the gamma function.

The mean and variance of a beta distribution with parameters α and β are

$$\mu = \frac{\alpha}{\alpha+\beta} \text{ and } \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

respectively.

For the uniform distribution on $(0,1)$, the mean and variance are

$$\mu = \frac{1}{1+1} = \frac{1}{2} \text{ and } \sigma^2 = \frac{(1)(1)}{(1+1)^2(1+1+1)} = \frac{1}{12},$$

respectively.

- **Note that the uniform distribution on (0, 1) is a beta distribution with parameters α = 1 and β = 1.**

Some Continuous Probability Distributions

Lognormal Distribution

- The distribution applies in cases where a natural log transformation results in a normal distribution.

The continuous random variable X has a **lognormal distribution** if the random variable $Y = \ln(X)$ has a normal distribution with mean μ and standard deviation σ. The resulting density function of X is

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

The mean and variance of the lognormal distribution are

$$\mu = e^{\mu+\sigma^2/2} \text{ and } \sigma^2 = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

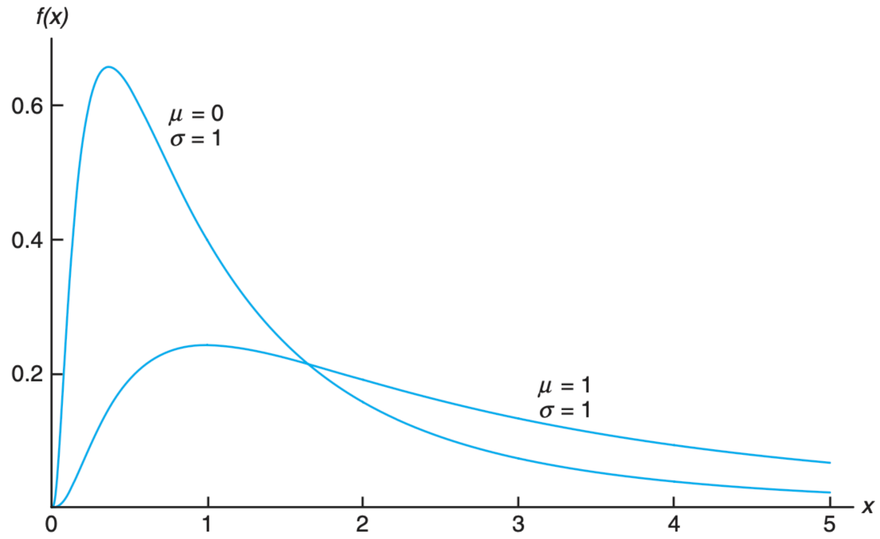


Figure 6.29: Lognormal distributions.

- **The distribution applies in cases where a natural log transformation results in a normal distribution.**
- **The cumulative distribution function is quite simple due to its relationship to the normal distribution.**

Some Continuous Probability Distributions

Weibull Distribution

- The distribution applies in cases where a natural log transformation results in a normal distribution.

The continuous random variable X has a **Weibull distribution**, with parameters α and β, if its density function is given by

$$f(x;\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & x > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$.

The mean and variance of the Weibull distribution are

$$\mu = \alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right) \text{ and } \sigma^2 = \alpha^{-2/\beta}\left\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\right\}.$$

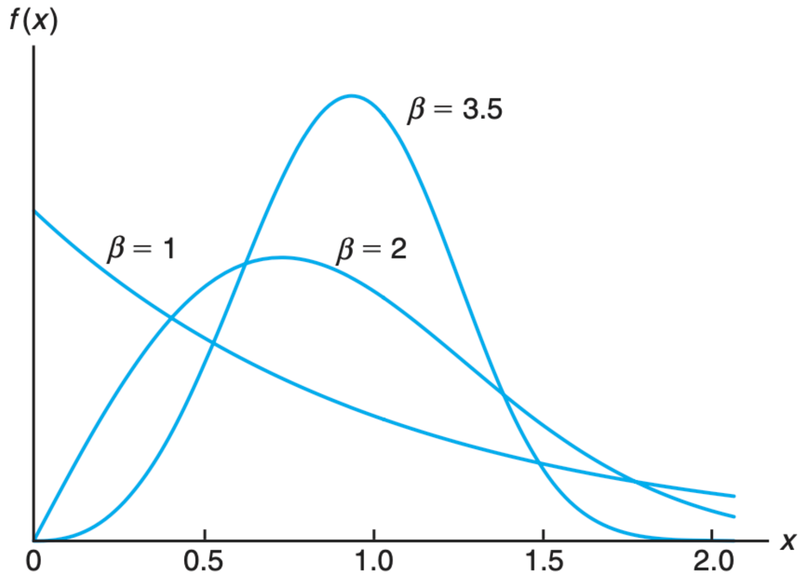


Figure 6.30: Weibull distributions ($\alpha = 1$).

- **The curves change considerably in shape for different values of the parameter β.**
- **If we let β = 1, the Weibull distribution reduces to the exponential distribution.**
- **For values of β > 1, the curves become somewhat bell shaped and resemble the normal curve but display some skewness.**

Functions of Random Variables

Transformations of Variables

Suppose that X is a **discrete** random variable with probability distribution $f(x)$. Let $Y = u(X)$ define a one-to-one transformation between the values of X and Y so that the equation $y = u(x)$ can be uniquely solved for x in terms of y, say $x = w(y)$. Then the probability distribution of Y is

$$g(y) = f[w(y)].$$

Suppose that X is a **continuous** random variable with probability distribution $f(x)$. Let $Y = u(X)$ define a one-to-one correspondence between the values of X and Y so that the equation $y = u(x)$ can be uniquely solved for x in terms of y, say $x = w(y)$. Then the probability distribution of Y is

$$g(y) = f[w(y)]|J|,$$

where $J = w'(y)$ and is called the **Jacobian** of the transformation.

- **To find the joint probability distribution of the random variables $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ when X_1 and X_2 are continuous and the transformation is one-to-one, we need an additional theorem.**

Suppose that X_1 and X_2 are **discrete** random variables with joint probability distribution $f(x_1, x_2)$. Let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ define a one-to-one transformation between the points (x_1, x_2) and (y_1, y_2) so that the equations

$$y_1 = u_1(x_1, x_2) \qquad \text{and} \qquad y_2 = u_2(x_1, x_2)$$

may be uniquely solved for x_1 and x_2 in terms of y_1 and y_2, say $x_1 = w_1(y_1, y_2)$ and $x_2 = w_2(y_1, y_2)$. Then the joint probability distribution of Y_1 and Y_2 is

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)].$$

Suppose that X_1 and X_2 are **continuous** random variables with joint probability distribution $f(x_1, x_2)$. Let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ define a one-to-one transformation between the points (x_1, x_2) and (y_1, y_2) so that the equations $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ may be uniquely solved for x_1 and x_2 in terms of y_1 and y_2, say $x_1 = w_1(y_l, y_2)$ and $x_2 = w_2(y_1, y_2)$. Then the joint probability distribution of Y_1 and Y_2 is

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|,$$

where the Jacobian is the 2×2 determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

and $\frac{\partial x_1}{\partial y_1}$ is simply the derivative of $x_1 = w_1(y_1, y_2)$ with respect to y_1 with y_2 held constant, referred to in calculus as the partial derivative of x_1 with respect to y_1. The other partial derivatives are defined in a similar manner.

Functions of Random Variables

Transformations of Variables

Example 7.4: Let X_1 and X_2 be two continuous random variables with joint probability distribution

$$f(x_1, x_2) = \begin{cases} 4x_1x_2, & 0 < x_1 < 1,\ 0 < x_2 < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the joint probability distribution of $Y_1 = X_1^2$ and $Y_2 = X_1X_2$.

***Solution*:** The inverse solutions of $y_1 = x_1^2$ and $y_2 = x_1x_2$ are $x_1 = \sqrt{y_1}$ and $x_2 = y_2/\sqrt{y_1}$, from which we obtain

$$J = \begin{vmatrix} 1/(2\sqrt{y_1}) & 0 \\ -y_2/2y_1^{3/2} & 1/\sqrt{y_1} \end{vmatrix} = \frac{1}{2y_1}.$$

To determine the set B of points in the y_1y_2 plane into which the set A of points in the x_1x_2 plane is mapped, we write

$$x_1 = \sqrt{y_1} \quad \text{and} \quad x_2 = y_2/\sqrt{y_1}.$$

$$g(y_1, y_2) = 4(\sqrt{y_1})\frac{y_2}{\sqrt{y_1}}\frac{1}{2y_1} = \begin{cases} \frac{2y_2}{y_1}, & y_2^2 < y_1 < 1,\ 0 < y_2 < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

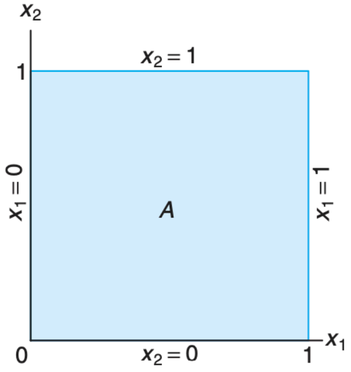
Suppose that X_1 and X_2 are **continuous** random variables with joint probability distribution $f(x_1, x_2)$. Let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ define a one-to-one transformation between the points (x_1, x_2) and (y_1, y_2) so that the equations $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ may be uniquely solved for x_1 and x_2 in terms of y_1 and y_2, say $x_1 = w_1(y_l, y_2)$ and $x_2 = w_2(y_1, y_2)$. Then the joint probability distribution of Y_1 and Y_2 is

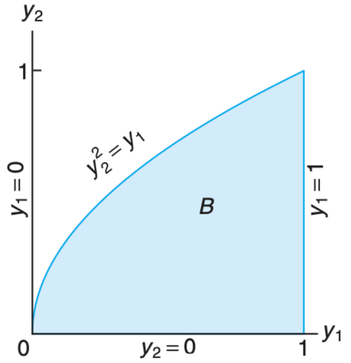
$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|,$$

where the Jacobian is the 2×2 determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

and $\frac{\partial x_1}{\partial y_1}$ is simply the derivative of $x_1 = w_1(y_1, y_2)$ with respect to y_1 with y_2 held constant, referred to in calculus as the partial derivative of x_1 with respect to y_1. The other partial derivatives are defined in a similar manner.





Fundamental Sampling Distributions and Data Descriptions

Populations and Samples:

A **population** consists of the totality of the observations with which we are concerned.

- The number of observations in the population is defined to be the size of the population.

A **sample** is a subset of a population.

- **Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be biased.**
- **To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a random sample in the sense that the observations are made independently and at random.**

Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

Fundamental Sampling Distributions and Data Descriptions

Some Important Statistics

- Statistic: Any function of the random variables constituting a random sample is called a statistic.

(a) Sample mean:

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

(b) Sample median:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd}, \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even}. \end{cases}$$

(c) The sample mode is the value of the sample that occurs most often.

(a) Sample variance:

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

(b) Sample standard deviation:

$$S = \sqrt{S^2},$$

where S^2 is the sample variance.

(c) Sample range:

$$R = X_{\max} - X_{\min}.$$

Fundamental Sampling Distributions and Data Descriptions

Sample Information

The probability distribution of a statistic is called a **sampling distribution.**

- The sampling distribution of a statistic depends on the distribution of the pop- ulation, the size of the samples, and the method of choosing the samples.

Sampling Distribution of Mean

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

The Central Limit Theorem

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

- **The sample size n = 30 is a guideline to use for the Central Limit Theorem**
- **The normal approximation for $\bar{X}$ will generally be good if n ≥ 30, provided the population distribution is not terribly skewed. If n < 30, the approximation is good only if the population is not too different from a normal distribution and, as stated above, if the population is known to be normal, the sampling distribution of $\bar{X}$ will follow a normal distribution exactly, no matter how small the size of the samples.**

Fundamental Sampling Distributions and Data Descriptions

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Inferences on the Population Mean

- One very important application of the Central Limit Theorem is the determination of reasonable values of the population mean μ.
- Topics such as hypothesis testing, estimation, quality control, and many others make use of the Central Limit Theorem.

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

***Solution*:** In this case, $\mu = 28$ and $\sigma = 3$. We need to calculate the probability $P(\bar{X} > 30)$ with $n = 40$. Since the time is measured on a continuous scale to the nearest minute, an $\bar{x}$ greater than 30 is equivalent to $\bar{x} \geq 30.5$. Hence,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes. An illustrative graph is shown in Figure 8.4.

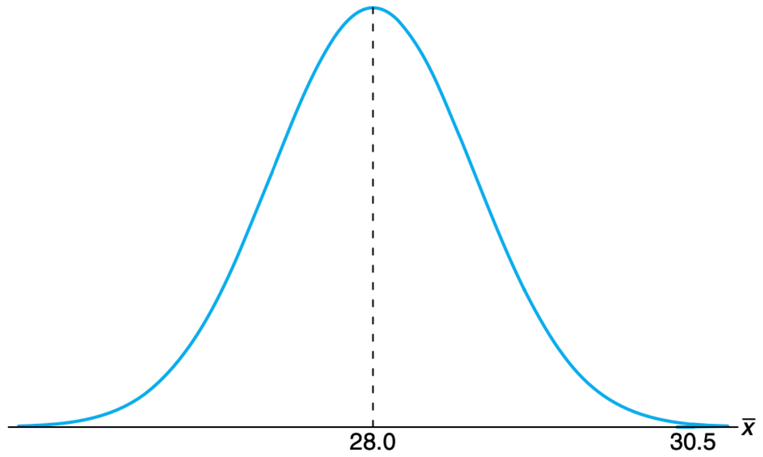


Figure 8.4: Area for Example 8.5.

Sampling Distribution of the Difference between Two Means

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately a standard normal variable.

Example 8.6: The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

If we use Theorem 8.3, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ will be approximately normal and will have a mean and standard deviation

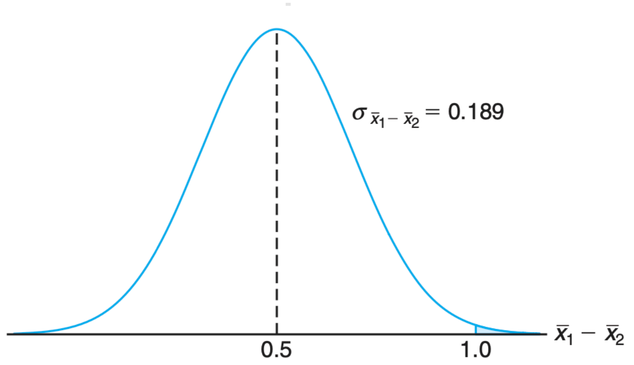
$$\mu_{\bar{X}_1-\bar{X}_2} = 6.5 - 6.0 = 0.5 \quad \text{and} \quad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189.$$

The probability that the mean lifetime for 36 tubes from manufacturer A will be at least 1 year longer than the mean lifetime for 49 tubes from manufacturer B is given by the area of the shaded region in Figure 8.6. Corresponding to the value $\bar{x}_1 - \bar{x}_2 = 1.0$, we find that

$$z = \frac{1.0 - 0.5}{0.189} = 2.65,$$

and hence

$$\begin{aligned} P(\bar{X}_1 - \bar{X}_2 \geq 1.0) &= P(Z > 2.65) = 1 - P(Z < 2.65) \\ &= 1 - 0.9960 = 0.0040. \end{aligned}$$



Fundamental Sampling Distributions and Data Descriptions

Sampling Distribution of S^2

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

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

has a chi-squared distribution with $v = n - 1$ degrees of freedom.

By the addition and subtraction of the sample mean $\bar{X}$, it is easy to see that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2 \\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) \\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2.\end{aligned}$$

Dividing each term of the equality by σ^2 and substituting $(n-1)S^2$ for $\sum_{i=1}^{n}(X_i-\bar{X})^2$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Now, according to Corollary 7.1 on page 222, we know that

$$\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$$

$$\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$$

has a χ^2-distribution with n degrees of freedom.

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

has a χ^2-distribution with n–1 degrees of freedom.

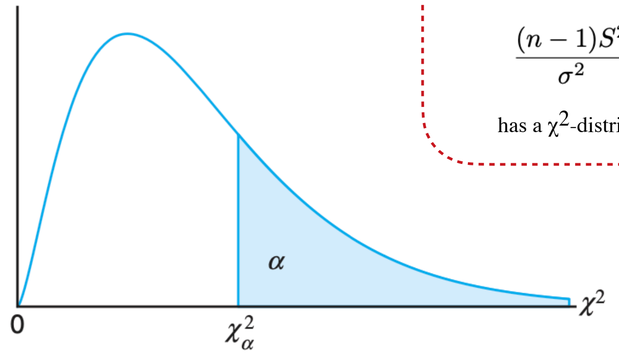


Figure 8.7: The chi-squared distribution.

- **Exactly 95% of a chi-squared distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$.**
- **There are n degrees of freedom, or independent pieces of information, in the random sample from the normal distribution. When the data (the values in the sample) are used to compute the mean, there is 1 less degree of freedom in the information used to estimate σ^2**

is a chi-squared random variable with n degrees of freedom. We have a chi-squared random variable with n degrees of freedom partitioned into two components. Note that in Section 6.7 we showed that a chi-squared distribution is a special case of a gamma distribution. The second term on the right-hand side is Z^2, which is a chi-squared random variable with 1 degree of freedom, and it turns out that $(n-1)S^2/\sigma^2$ is a chi-squared random variable with n – 1 degree of freedom. We formalize this in the following theorem.

Fundamental Sampling Distributions and Data Descriptions

t-Distribution

Let Z be a standard normal random variable and V a chi-squared random variable with v degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where

$$T = \frac{Z}{\sqrt{V/v}},$$

is given by the density function

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

This is known as the ***t*-distribution** with v degrees of freedom.

Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

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

Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $v = n - 1$ degrees of freedom.

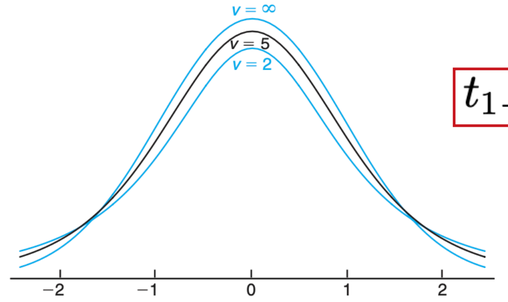


Figure 8.8: The t-distribution curves for $v = 2, 5,$ and ∞.

$$t_{1-\alpha} = -t_{\alpha}$$

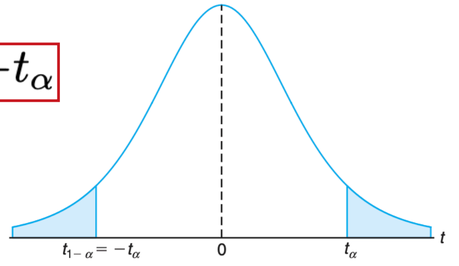


Figure 8.9: Symmetry property (about 0) of the t-distribution.

- **Exactly 95% of the values of a t-distribution with v = n – 1 degrees of freedom lie between $-t_{0.025}$ and $t_{0.025}$.**
- **The shortest possible interval is obtained by choosing t-values that leave exactly the same area in the two tails of our distribution.**
- **A t-value that falls below $-t_{0.025}$ or above $t_{0.025}$ would tend to make us believe either that a very rare event has taken place or that our assumption about μ is in error.**
- **The t-distribution is used extensively in problems that deal with inference about the population mean (as illustrated in Example 8.11) or in problems that involve comparative samples (i.e., in cases where one is trying to determine if means from two samples are significantly different).**

Fundamental Sampling Distributions and Data Descriptions

F-Distribution

Let U and V be two independent random variables having chi-squared distributions with v_1 and v_2 degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density function

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

This is known as the ***F*-distribution** with v_1 and v_2 degrees of freedom (d.f.).

Writing $f_\alpha(v_1, v_2)$ for f_α with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

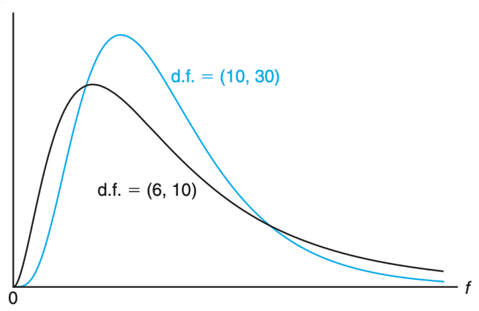


Figure 8.11: Typical F-distributions.

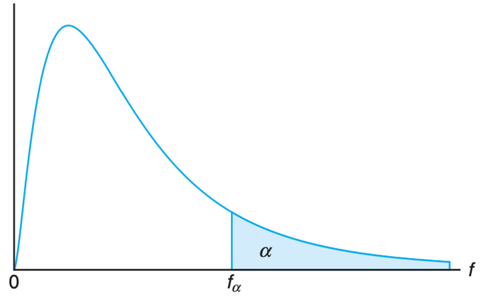


Figure 8.12: Illustration of the f_α for the F-distribution.

- **The F-distribution finds enormous application in comparing sample variances.**
- **The statistic F is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom.**
- **However, the density function will not be used and is given only for completeness. The curve of the F-distribution depends not only on the two parameters v_1 and v_2 but also on the order in which we state them.**
- **The F -distribution can also be applied to many other types of problems involving sam- ple variances. In fact, the F-distribution is called the variance ratio distribution.**

Fundamental Sampling Distributions and Data Descriptions

F-Distribution

- Case study: we wish to determine if the population means are equivalent.

Paint	Sample Mean	Sample Variance	Sample Size
A	$\bar{X}_A = 4.5$	$s_A^2 = 0.20$	10
B	$\bar{X}_B = 5.5$	$s_B^2 = 0.14$	10
C	$\bar{X}_C = 6.5$	$s_C^2 = 0.11$	10

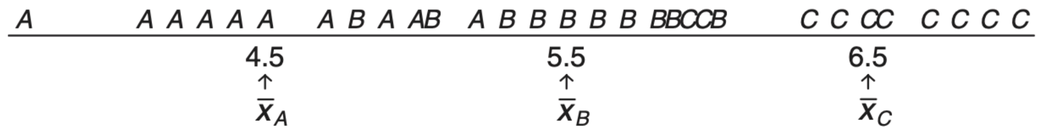


Figure 8.13: Data from three distinct samples.

- **Whether these sample averages could have occurred by chance depends on the variability within samples.**

(1) Variability within samples (between observations in distinct samples)
(2) Variability between samples (between sample averages)

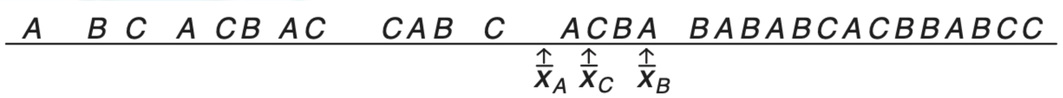


Figure 8.14: Data that easily could have come from the same population.

The sources of variability in (1) and (2) above generate important ratios of sample variances, and ratios are used in conjunction with the F -distribution. The general procedure involved is called analysis of variance. It is interesting that in the paint example described here, we are dealing with inferences on three population means, but two sources of variability are used. We will not supply details here, but in Chapters 13 through 15 we make extensive use of analysis of variance, and, of course, the F-distribution plays an important role.

Fundamental Sampling Distributions and Data Descriptions

Quantile and Probability Plots

A **quantile** of a sample, $q(f)$, is a value for which a specified fraction f of the data values is less than or equal to $q(f)$.

- The purpose of the quantile plot is to depict, in sample form, the cumulative distribution function
- A quantile plot simply plots the data values on the vertical axis against an empirical assessment of the fraction of observations exceeded by the data value.

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}},$$

where i is the order of the observations when they are ranked from low to high. In other words, if we denote the ranked observations as

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)},$$

then the quantile plot depicts a plot of $y_{(i)}$ against f_i.

- **Quantile plots are useful in detection of distribution types.**
- **There are also situations in both model building and design of experiments in which the plots are used to detect important model terms or effects that are active.**
- **They are used to determine whether or not the underlying assumptions made by the scientist or engineer in building the model are reasonable.**

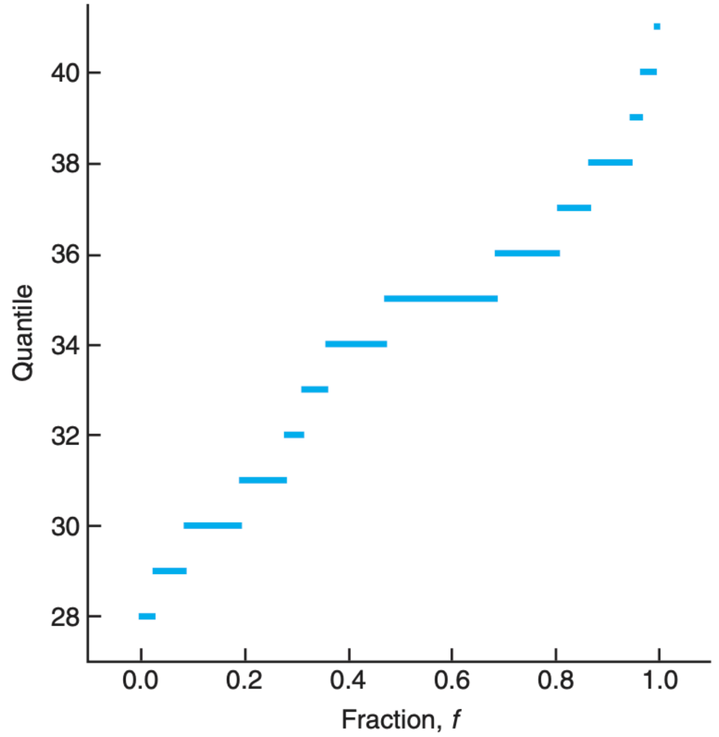


Figure 8.15: Quantile plot for paint data.

Fundamental Sampling Distributions and Data Descriptions

Normal Quantile-Quantile Plot

The **normal quantile-quantile plot** is a plot of $y_{(i)}$ (ordered observations) against $q_{0,1}(f_i)$, where $f_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$.

- The methodology involves a plot of the empirical quantiles recently discussed against the corresponding quantile of the normal distribution.
- The expression for a quantile of an N(μ,σ) random variable is very complicated.

$$q_{\mu,\sigma}(f) = \mu + \sigma\{4.91[f^{0.14} - (1-f)^{0.14}]\}.$$

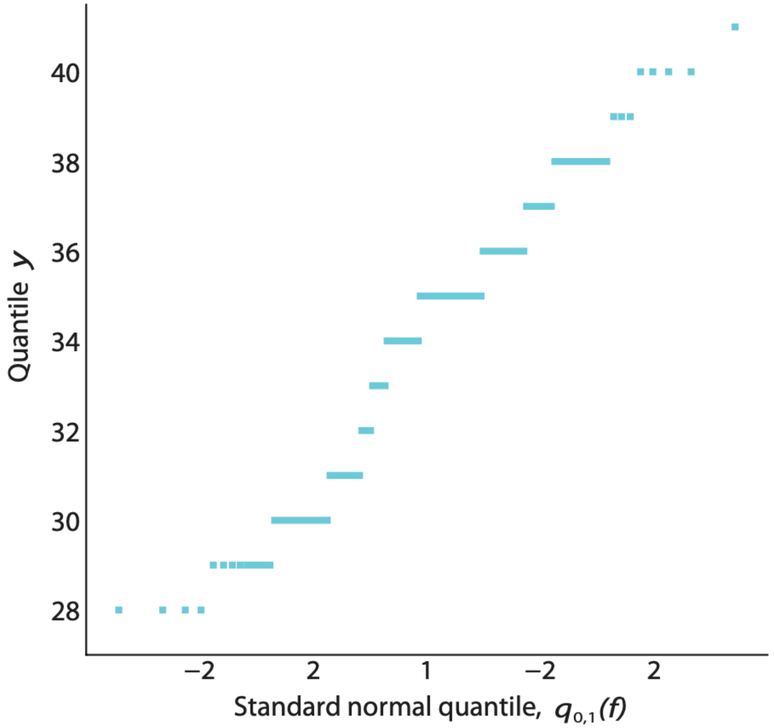


Figure 8.16: Normal quantile-quantile plot for paint data.

- **The normal quantile-quantile plot takes advantage of what is known about the quantiles of the normal distribution.**
- **The methodology involves a plot of the empirical quantiles recently discussed against the corresponding quantile of the normal distribution.**

Applied Statistics

Jiebo Song

songjiebo@bimsa.cn

北京雁栖湖
应用数学研究院
Beijing Institute
of Mathematical
Sciences and Applications

Fundamental Sampling Distribution

* Continuous *Sampling *Statistics