STATISTICAL INFERENCES (2cr)
Chapter 8 Sampling Distributions & Data Descriptions

Zhong Guan

Math, IUSB

Outline

Sampling distribution of s_1^2/s_2^2

- Let $x_{11}, \ldots, x_{1n_1}$ be a sample from normal population with variance σ_1^2,
and $x_{21}, \ldots, x_{2n_2}$ be a sample from normal population with variance σ_2^2.
- Assume the two samples are independent and have sample variances s_1^2 and s_2^2, respectively.
- The distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is F-distribution with numerator degrees of freedom $\nu_1 = n_1 - 1$ and denominator degrees of freedom $\nu_2 = n_2 - 1$.
- Generally, if U and V are independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively, then $F = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution with numerator and denominator degrees of freedom ν_1 and ν_2.

Sampling distribution of s_1^2/s_2^2

- Let $x_{11}, \ldots, x_{1n_1}$ be a sample from normal population with variance σ_1^2,
  and $x_{21}, \ldots, x_{2n_2}$ be a sample from normal population with variance σ_2^2.
- Assume the two samples are independent and have sample variances s_1^2 and s_2^2, respectively.
- The distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is F-distribution with numerator degrees of freedom $\nu_1 = n_1 - 1$ and denominator degrees of freedom $\nu_2 = n_2 - 1$.
- Generally, if U and V are independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively, then $F = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution with numerator and denominator degrees of freedom ν_1 and ν_2.

Sampling distribution of s_1^2/s_2^2

- Let $x_{11}, \ldots, x_{1n_1}$ be a sample from normal population with variance σ_1^2,
  and $x_{21}, \ldots, x_{2n_2}$ be a sample from normal population with variance σ_2^2.
- Assume the two samples are independent and have sample variances s_1^2 and s_2^2, respectively.
- The distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is F-distribution with numerator degrees of freedom $\nu_1 = n_1 - 1$ and denominator degrees of freedom $\nu_2 = n_2 - 1$.
- Generally, if U and V are independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively, then $F = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution with numerator and denominator degrees of freedom ν_1 and ν_2.

Sampling distribution of s_1^2/s_2^2

- Let $x_{11}, \ldots, x_{1n_1}$ be a sample from normal population with variance σ_1^2,
  and $x_{21}, \ldots, x_{2n_2}$ be a sample from normal population with variance σ_2^2.
- Assume the two samples are independent and have sample variances s_1^2 and s_2^2, respectively.
- The distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is F-distribution with numerator degrees of freedom $\nu_1 = n_1 - 1$ and denominator degrees of freedom $\nu_2 = n_2 - 1$.
- Generally, if U and V are independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively, then $F = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution with numerator and denominator degrees of freedom ν_1 and ν_2.

Sampling distribution of s_1^2/s_2^2

- Let $x_{11}, \ldots, x_{1n_1}$ be a sample from normal population with variance σ_1^2,
  and $x_{21}, \ldots, x_{2n_2}$ be a sample from normal population with variance σ_2^2.
- Assume the two samples are independent and have sample variances s_1^2 and s_2^2, respectively.
- The distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is F-distribution with numerator degrees of freedom $\nu_1 = n_1 - 1$ and denominator degrees of freedom $\nu_2 = n_2 - 1$.
- Generally, if U and V are independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively, then $F = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution with numerator and denominator degrees of freedom ν_1 and ν_2.

The F Critical Value (Table A.6)

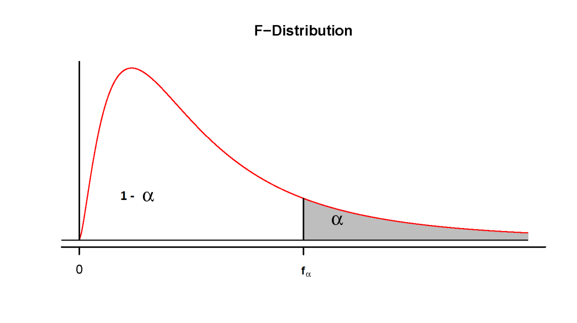


Table A.6 gives $f_\alpha(\nu_1, \nu_2)$ for $\alpha = .05, .01$.
For example, $\nu_1 = 5$, $\nu_2 = 7$, $f_{.05}(5,7) = 3.97$ by Table A.6.

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$$

The *F* Critical Value (Table A.6)

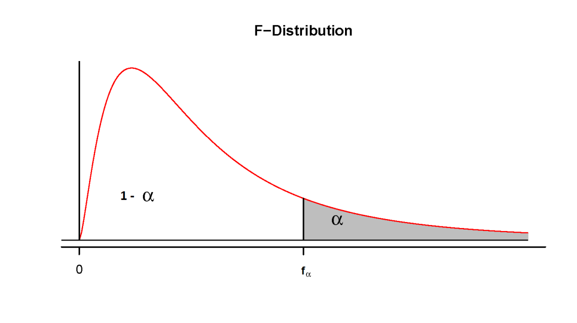
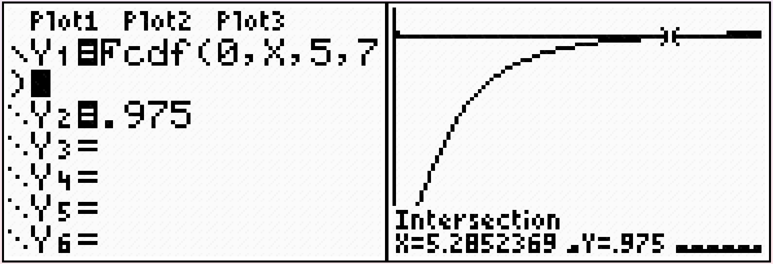


Table A.6 gives $f_\alpha(\nu_1, \nu_2)$ for $\alpha = .05, .01$.
For example, $\nu_1 = 5$, $\nu_2 = 7$, $f_{.05}(5, 7) = 3.97$ by Table A.6.

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$$

The *F* Critical Value using Technology

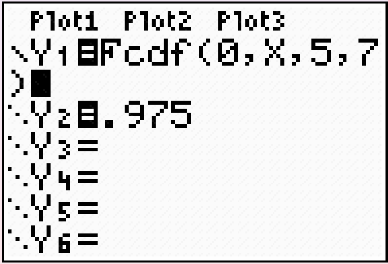
- In Excel: $f_{.025}(5,7)$=F.INV.RT(0.025,5,7)=5.285236852, or =F.INV(1-0.025,5,7)=5.285236852.
- In TI-8x: Graph Y1=Fcdf(0, X, 5,7) and Y2=1-.025=0.975, the *x*-coordinate of the intersection is $f_{.025}(5,7)$=5.2852369.

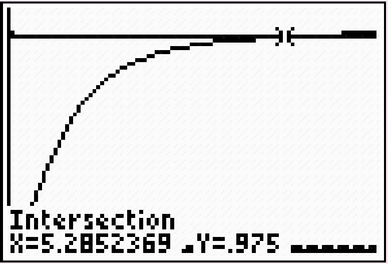


- In R: $f_{.025}$= "qf(.975,5,7)=5.285237".

The *F* Critical Value using Technology

- In Excel: $f_{.025}(5,7)$=F.INV.RT(0.025,5,7)=5.285236852, or =F.INV(1-0.025,5,7)=5.285236852.
- In TI-8x: Graph Y1=Fcdf(0, X, 5,7) and Y2=1-.025=0.975, the x-coordinate of the intersection is $f_{.025}(5,7)$=5.2852369.

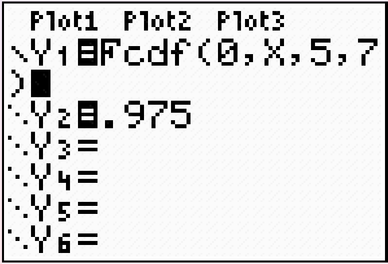


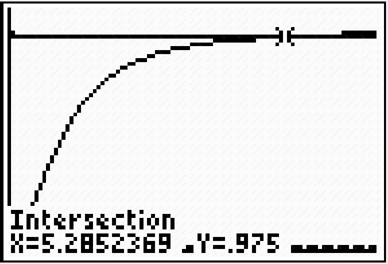


- In R: $f_{.025}$= “qf(.975,5,7)=5.285237”.

The F Critical Value using Technology

- In Excel: $f_{.025}(5,7)$=F.INV.RT(0.025,5,7)=5.285236852, or =F.INV(1-0.025,5,7)=5.285236852.
- In TI-8x: Graph Y1=Fcdf(0, X, 5,7) and Y2=1-.025=0.975, the x-coordinate of the intersection is $f_{.025}(5,7)$=5.2852369.





- In R: $f_{.025}$= “qf(.975,5,7)=5.285237”.

Definition

Sample Quantile: A quantile of a sample is a value, $q(p)$, for which a specific proportion p of the data values is less than or equal to $q(p)$.

Specifically, Let $x_1, \ldots, x_n$ be the sample data. Then $q(p)$ is the smallest data value such that the proportion of the data values less than or equal to $q(p)$ is at least p.

Definition

Sample Quantile: A quantile of a sample is a value, $q(p)$, for which a specific proportion p of the data values is less than or equal to $q(p)$.
Specifically, Let $x_1, \ldots, x_n$ be the sample data. Then $q(p)$ is the smallest data value such that the proportion of the data values less than or equal to $q(p)$ is at least p.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.
MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.
MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.
MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.
MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.

MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.
MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Example.

Example 1.: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58

(a) Find and interpret the 15% quantile $q(0.15)$;

(b) Find and interpret the 75% quantile $q(0.75)$;

Example.

Example 1.: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58

(a) Find and interpret the 15% quantile $q(0.15)$;

(b) Find and interpret the 75% quantile $q(0.75)$;

Definition

Population Quantile: A quantile of a population with pmf/pdf $f(x)$ is a value, $q_f(p)$, for which the probability that any population data value is less than or equal to $q_f(p)$ about p, that is

$$P(X \leqslant q_f(p)) \approx p, \quad q_f(p) = \min\{x : P(X \leqslant x) \geqslant p\}.$$

For discret distribution

$$\sum_{x \leqslant q_f(p)} f(x) \approx p$$

For continuous distribution

$$F(q_f(p)) = \int_{-\infty}^{q_f(p)} f(x)dx = p, \quad \text{i.e.} \quad q_f(p) = F^{-1}(p)$$

Definition

Population Quantile: A quantile of a population with pmf/pdf $f(x)$ is a value, $q_f(p)$, for which the probability that any population data value is less than or equal to $q_f(p)$ about p, that is

$$P(X \leqslant q_f(p)) \approx p, \quad q_f(p) = \min\{x : P(X \leqslant x) \geqslant p\}.$$

For discret distribution

$$\sum_{x \leqslant q_f(p)} f(x) \approx p$$

For continuous distribution

$$F(q_f(p)) = \int_{-\infty}^{q_f(p)} f(x)dx = p, \quad \text{i.e.} \quad q_f(p) = F^{-1}(p)$$

Definition

Population Quantile: A quantile of a population with pmf/pdf $f(x)$ is a value, $q_f(p)$, for which the probability that any population data value is less than or equal to $q_f(p)$ about p, that is

$$P(X \leqslant q_f(p)) \approx p, \quad q_f(p) = \min\{x : P(X \leqslant x) \geqslant p\}.$$

For discret distribution

$$\sum_{x \leqslant q_f(p)} f(x) \approx p$$

For continuous distribution

$$F(q_f(p)) = \int_{-\infty}^{q_f(p)} f(x)dx = p, \quad \text{i.e.} \quad q_f(p) = F^{-1}(p)$$

Finding Population Quantiles $q_f(p)$

Example 2. Let the population distribution of X have pmf

$$f(x) = \frac{x}{10}, \quad x = 1, 2, 3, 4$$

Find $q_f(0.1)$ and $q_f(0.6)$.

Example 3. Let the population distribution of X have pdf

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

Find $q_f(p)$ for $0 < p < 1$.

Finding Population Quantiles $q_f(p)$

Example 2. Let the population distribution of X have pmf

$$f(x) = \frac{x}{10}, \quad x = 1, 2, 3, 4$$

Find $q_f(0.1)$ and $q_f(0.6)$.
Example 3. Let the population distribution of X have pdf

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

Find $q_f(p)$ for $0 < p < 1$.

Definition

Quantile-Quantile(Q-Q) Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$.

Usage If the points of the q-q plot are close to a straight line, then the data is likely from the distribution f.

Definition

Quantile-Quantile(Q-Q) Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$.

Usage If the points of the q-q plot are close to a straight line, then the data is likely from the distribution f.

Definition

Exponential Q-Q Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$, and f is the exponential distribution with mean =1, i.e., $q_f(p) = -\ln(1-p)$, $0 < p < 1$.

Usage If the points of the q-q plot are close to a straight line with y-intercept 0, then the data is likely from an exponential distribution with mean being estimated by the slope.

Example 4. Construct an exponential q-q lot for the data: 1.094, 2.630, 0.882, 1.885, 0.721, 1.290, 0.019.

Definition

Exponential Q-Q Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$, and f is the exponential distribution with mean =1, i.e., $q_f(p) = -\ln(1-p)$, $0 < p < 1$.

Usage If the points of the q-q plot are close to a straight line with y-intercept 0, then the data is likely from an exponential distribution with mean being estimated by the slope.

Example 4. Construct an exponential q-q lot for the data: 1.094, 2.630, 0.882, 1.885, 0.721, 1.290, 0.019.

Definition

Exponential Q-Q Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$, and f is the exponential distribution with mean =1, i.e., $q_f(p) = -\ln(1-p)$, $0 < p < 1$.

Usage If the points of the q-q plot are close to a straight line with y-intercept 0, then the data is likely from an exponential distribution with mean being estimated by the slope.

Example 4. Construct an exponential q-q lot for the data: 1.094, 2.630, 0.882, 1.885, 0.721, 1.290, 0.019.

Definition

Normal q-q plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where f is the standard normal distribution $N(0, 1)$ and $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$. $q_f(p) = \Phi^{-1}(p)$=invNorm(p).

Usage If the points of the q-q plot are close to a straight line, then the data is likely from a normal distribution $N(\mu, \sigma^2)$. The slope is an estimate of σ and y-intercept is an estimate of μ.

Example 5. Construct a normal q-q lot for the data:

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75	6.66
6.66	6.64	6.76	6.73	6.80	6.72	6.76	6.76	6.68
6.66	6.62	6.72	6.76	6.70	6.78	6.76	6.67	6.70
6.72	6.74	6.81	6.79	6.78	6.66	6.76	6.76	6.72

Definition

Normal q-q plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where f is the standard normal distribution $N(0,1)$ and $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$. $q_f(p) = \Phi^{-1}(p)$=invNorm(p).

Usage If the points of the q-q plot are close to a straight line, then the data is likely from a normal distribution $N(\mu, \sigma^2)$. The slope is an estimate of σ and y-intercept is an estimate of μ.

Example 5. Construct a normal q-q lot for the data:

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75	6.66
6.66	6.64	6.76	6.73	6.80	6.72	6.76	6.76	6.68
6.66	6.62	6.72	6.76	6.70	6.78	6.76	6.67	6.70
6.72	6.74	6.81	6.79	6.78	6.66	6.76	6.76	6.72

Definition

Normal q-q plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where f is the standard normal distribution $N(0,1)$ and $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$. $q_f(p) = \Phi^{-1}(p)$=invNorm(p).

Usage If the points of the q-q plot are close to a straight line, then the data is likely from a normal distribution $N(\mu, \sigma^2)$. The slope is an estimate of σ and y-intercept is an estimate of μ.

Example 5. Construct a normal q-q lot for the data:

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75	6.66
6.66	6.64	6.76	6.73	6.80	6.72	6.76	6.76	6.68
6.66	6.62	6.72	6.76	6.70	6.78	6.76	6.67	6.70
6.72	6.74	6.81	6.79	6.78	6.66	6.76	6.76	6.72