Linear Statistical Modeling Methods with SAS

Statistical Inferences about Population Means

Xuemao Zhang
East Stroudsburg University

January 29, 2024

Outline

- Sampling Distributions and Point Estimators
- Confidence Interval Estimations of a Population Mean
- Hypothesis Testing

Sampling Distributions

Random sample

The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the population with pdf/pmf $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called **independent and identically distributed** random variables with pdf or pmf $f(x)$. This is commonly abbreviated to iid random variables.

Remark. A sample drawn from a finite population without replacement does not satisfy all conditions of the above definition. The random variables $X_1, \ldots, X_n$ are not mutually independent.

Statistic and sampling distribution

Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $X_1, \ldots, X_n$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution** of Y.

The Central Limit Theorem

Convergence in distribution A sequence of random variables, $X_1, X_2, \ldots$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{x_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be a sequence of iid random variables. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Define $\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$. Let $G_n(x)$ denote the cdf of $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$. Then, for any $x, -\infty < x < \infty$,

$$\lim_{n\to\infty} P(G_n(x) \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

That is, $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ has a limiting standard normal distribution.

The Central Limit Theorem

- A simulation study: Consider repeating this process: Roll a balanced die 5 times. Find the mean of the results.
  Let X be the random variable of the results of rolling the die. We conduct a Monte Carlo simulation of repeating the process (of rolling the die 5 times) 10,000 times to study the sampling distribution of the sample mean.

```
proc iml;

iter = 10000;      /*iteration times*/
n=5;              /*sample size*/
means=j(iter, 1, 0);   /* The sample means in the form of iter*1 matrix*/
Min = 1; Max = 6;
Do i=1 to iter;
 x=j(n,1,0);   /* The sample data in the form of n*1 matrix*/
  do j = 1 to n;
  x[j]=min + floor((1+Max-Min)*rand("Uniform")); /* random integer values i
  end;
means[i]=mean(x);
End;
```

The Central Limit Theorem

```
xbar_bar=mean(means);
print xbar_bar;
title "Histogram of the sample means";
call Histogram(means);
quit;
```

The Central Limit Theorem

- We then compare the simulation results with the theoretical results by using the probability distribution of X. Since the die is balanced, we have

  $$P(x) = P(X = x) = 1/6, x = 1, 2, 3, 4, 5, 6.$$

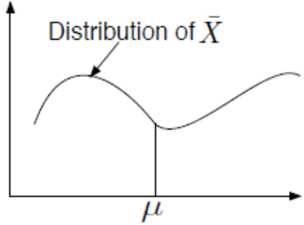
  Thus,

  $$\mu = \sum xP(x) = 21/6 = 3.5$$

The population mean is 3.5; the mean of the 10,000 trials in the simulation was 3.50872. If continued indefinitely, the sample mean will be 3.5. Also, notice the distribution is "approximately normal."

Point Estimators

In evaluating the value of parameters, the estimation of a parameter is essential as it is often difficult or impossible to get the exact parameter from a population.



Definition. An **estimator** is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.

Remark. An estimator is a **function** of the sample $X_1, \ldots, X_n$. For example, the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$.

Confidence Intervals

If we are trying to estimate a parameter θ, the point estimate $\widehat{\theta}$, combined with some idea of how accurate $\widehat{\theta}$ is, give a range of values that is likely to include the true value of θ. Such an interval is called Confidence Interval.

Definition 1. A $1-\alpha$ level two-sided confidence interval of θ is a *random interval* $[\widehat{\theta}_L, \widehat{\theta}_U]$ such that

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1-\alpha,$$

where $1-\alpha$ is called *confidence level* or *confidence coefficient*.

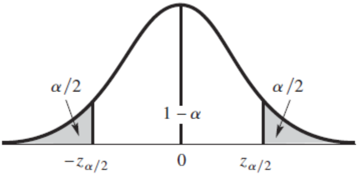
For Large Sample sizes, we know that

$$Z = \frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} \text{ (Pivotal quantity)}$$

is approximately standard normal.

Large-Sample Confidence Intervals

Then we can use Z to derive a two-sided $1-\alpha$ level confidence interval $[\widehat{\theta}_L, \widehat{\theta}_U]$ of θ.



Distribution of *Z*

Recall. The Z **critical value** with right tail area α is a point, denoted by Z_α, such that $P(Z \geq Z_\alpha) = \alpha$.

Now,

$$\begin{aligned} 1-\alpha =& P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) \approx P(-Z_{\alpha/2} \leq \frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}} \leq Z_{\alpha/2}) \\ =& P(\widehat{\theta} - Z_{\alpha/2}\sigma_{\widehat{\theta}} \leq \theta \leq \widehat{\theta} + Z_{\alpha/2}\sigma_{\widehat{\theta}}). \end{aligned}$$

That is, the two-sided $1-\alpha$ level confidence interval of θ is

$$\boxed{[\widehat{\theta} - Z_{\alpha/2}\sigma_{\widehat{\theta}}, \widehat{\theta} + Z_{\alpha/2}\sigma_{\widehat{\theta}}]}$$

Large-Sample Confidence Intervals

A two-sided $1-\alpha$ level confidence interval of a population mean μ is

$$\boxed{[\overline{X} - Z_{\alpha/2}\sigma_{\overline{X}}, \overline{X} + Z_{\alpha/2}\sigma_{\overline{X}}]},$$

where $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

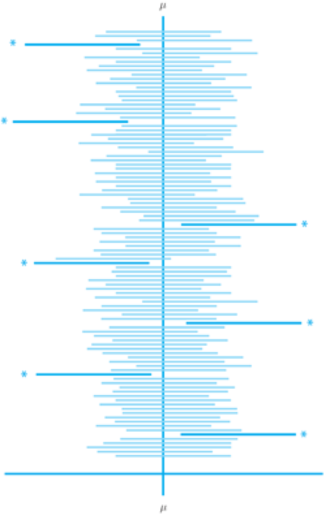
Remark.

1. The accuracy of the estimate is determined by $\frac{Z_{\alpha/2}\sigma}{\sqrt{n}}$, and gets better when:
   - $Z_{\alpha/2}$ gets smaller. To get a smaller interval is to take lower confidence.
   - Better estimates are obtained in populations with smaller inherent variability.
   - When the sample size n is larger.
2. $\sigma_{\widehat{\theta}}$ is estimated by $S/\sqrt{n}$.

Interpreting a Confidence Level

1. A $1-\alpha$ confidence interval is a random interval. A correct interpretation of $100(1-\alpha)\%$ confidence relies on the long-run relative frequency interpretation of probability in repeated sampling. That is, in the long run $100(1-\alpha)\%$ of our computed CIs will contain the true parameter, which is an unknown constant.
2. When we obtain a $100(1-\alpha)\%$ confidence interval from a sample, all randomness disappears; the calculated interval is not a random interval, and we should not use probability language to interpret the interval. See the illustration graph in the next slide.

Interpreting a Confidence Level



One hundred 95% CIs (asterisks identify intervals that do not include μ).

Other Types of Confidence Intervals

Denote the parameter of interest by θ.

Definition 2. A $1-\alpha$ level right-sided confidence interval $[\widehat{\theta}_L, \infty)$ is a a *random interval* such that

$$P(\theta \geq \widehat{\theta}_L) = 1-\alpha.$$

It can be used for the test of $H_0 : \theta = \theta_0$ against $H_A : \theta > \theta_0$

Definition 3. A $1-\alpha$ level left-sided confidence interval $(-\infty, \widehat{\theta}_U)$ is a a *random interval* such that

$$P(\theta \leq \widehat{\theta}_U) = 1-\alpha.$$

It can be used for the test of $H_0 : \theta = \theta_0$ against $H_A : \theta < \theta_0$

Small-Sample Confidence Intervals for μ

THEOREM. When $\overline{Y}$ is the mean of a random sample of size n from a normal distribution with mean μ, the rv

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

has a t-distribution with $n-1$ degrees of freedom (df).

Remark. The t-distribution of T is robust to small or even moderate departures from normality unless the sample size n is quite small.

THEOREM. Let $Y_1, \ldots, Y_n$ be a random sample from a normal population with mean μ.

(1) A two-sided $1-\alpha$ CI of μ is $\overline{Y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$.

(2) A one-sided $1-\alpha$ CI of μ is $[\overline{Y} - t_{\alpha}\left(\frac{S}{\sqrt{n}}\right), \infty)$.

(3) A one-sided $1-\alpha$ CI of μ is $(-\infty, \overline{Y} + t_{\alpha}\left(\frac{S}{\sqrt{n}}\right)]$.

where $t_{\alpha/2}$ or t_{α} is determined from the t distribution with $df = n-1$.

Example

Example. A manufacturer of gunpowder has developed a new powder, which was tested in eight shells. The resulting muzzle velocities, in feet per second, were as follows: 3005, 2925, 2935, 2965, 2995, 3005, 2937, 2905.

Assume that muzzle velocities are approximately normally distributed.

1. Find a 95% confidence interval for the true average velocity μ for shells of this type if the population standard deviation is known $\sigma = 39$.
2. Find a 95% confidence interval for the true average velocity μ for shells of this type if the population standard deviation is unknown.

Elements of a Statistical Test

One of the most important applications of statistics is to make inferences about the population from a random sample selected from the population.

Definition. A *hypothesis* is a statement about a population parameter.

The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.

Definition. The two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_a (or H_1), respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0 : \theta \in \Omega_0$ and $H_a : \theta \in \Omega_0^c$, where Ω_0 is some subset of the parameter space and Ω_0^c is its complement. There are three types of hypothesis tests

(1) $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.
(2) $H_0 : \theta \geq \theta_0$ versus $H_a : \theta < \theta_0$.
(3) $H_0 : \theta \leq \theta_0$ versus $H_a : \theta > \theta_0$.

Elements of a Statistical Test

Remark. The alternative H_a is the hypothesis that the researcher wishes to support. If we are conducting a study and want to use a hypothesis test to support our claim, the claim must be worded so that it becomes the alternative.

Definition. A *hypothesis testing procedure* or *hypothesis test* is a rule that specifies:

(i.) For which sample values H_0 is rejected and H_a is accepted as true.
(ii.) For which sample values the decision is made to NOT to reject H_0.

The subset of the sample space for which H_0 will be rejected is called the *rejection region* or *critical region*, denoted by RR. The complement of the rejection region is called the *acceptance region*.

The Elements of a Statistical Test.

(1) Null hypothesis, H_0
(2) Alternative hypothesis, H_a
(3) Test statistic
(4) Rejection region (RR)

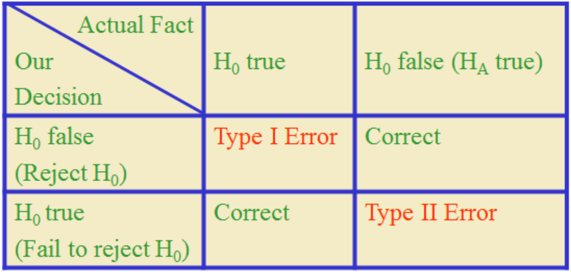
Elements of a Statistical Test

Remark. Finding a good rejection region for a statistical test is an interesting problem that merits further attention.

For any fixed rejection region, two types of errors can be made in reaching a decision.

Definition. A type I error is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted by α. α is called the **level** or **significance level** of the test. Some typical significance level is $\alpha = 0.1$, $\alpha = 0.05$ or $\alpha = 0.01$.

A type II error is made if H_0 is NOT rejected when H_a is true. The probability of a type II error is denoted by β.



Actual Fact Our Decision	H_0 true	H_0 false (H_A true)
H_0 false (Reject H_0)	Type I Error	Correct
H_0 true (Fail to reject H_0)	Correct	Type II Error

Elements of a Statistical Test

General Steps for Testing a Hypothesis.

(1) State the null hypothesis H_0 and the alternative hypothesis H_a.

(2) Pick a test statistic, a value calculated from the sample data that you will base your decision on.

(3) Choose a level of significance α and find the critical points showing the boundary of the rejection region. Determine if H_0 can be rejected.

(4) Make a decision, a statement that uses simple nontechnical wording that addresses the original claim.

Large Sample Tests

We summarize the hypothesis testing procedure based on an estimator $\widehat{\theta}$ that has an (approximately) normal distribution with mean θ and standard deviation $\sigma_{\widehat{\theta}}$.

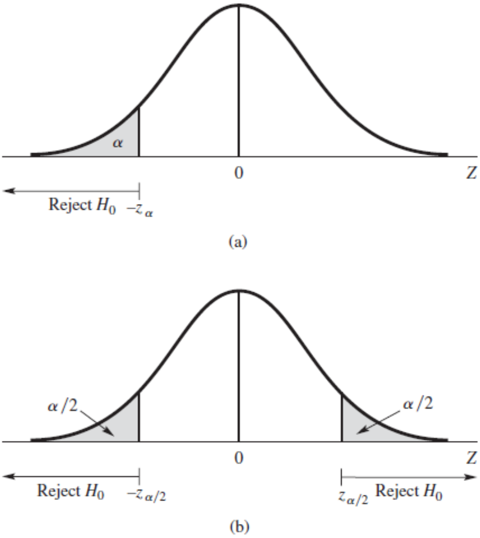
Large-Sample α-Level Hypothesis Tests.

$$\boxed{\begin{array}{l} H_0 : \theta = \theta_0 \\ H_a : \begin{cases} \theta > \theta_0, & \text{upper-tail alternative;} \\ \theta < \theta_0, & \text{lower-tail alternative;} \\ \theta \neq \theta_0, & \text{two-tailed alternative.} \end{cases} \\ \text{Test Statistic: } Z_0 = \dfrac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} \\ \text{Rejection Region: } RR = \begin{cases} \{Z : Z \geq Z_\alpha\}, & \text{upper-tail RR;} \\ \{Z : Z \leq -Z_\alpha\}, & \text{lower-tail RR;} \\ \{Z : |Z| \geq Z_{\alpha/2}\}, & \text{two-tailed RR.} \end{cases} \end{array}}$$

Large Sample Tests

Large-Sample α-Level Hypothesis Tests.

FIGURE 10.4 Rejection regions for testing $H_0 : \theta = \theta_0$ versus (a) $H_a : \theta < \theta_0$ and (b) $H_a : \theta \neq \theta_0$, based on $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$



One Sample t-Test

One Sample t-Test. Let $Y_1, \ldots, Y_n$ be a random sample from a population with approximate $\{$**Normal**$\}$ **distribution.**

$H_0 : \mu = \mu_0$

$$H_a : \begin{cases} \mu > \mu_0, & \textbf{upper-tail alternative;} \\ \mu < \mu_0, & \textbf{lower-tail alternative;} \\ \mu \neq \mu_0, & \textbf{two-tailed alternative.} \end{cases}$$

Test Statistic: $t_0 = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}}$

Rejection Region: $RR = \begin{cases} \{t : t \geq t_\alpha\}, & \textbf{upper-tail RR;} \\ \{t : t \leq -t_\alpha\}, & \textbf{lower-tail RR;} \\ \{t : |t| \geq t_{\alpha/2}\}, & \textbf{two-tailed RR.} \end{cases}$

where the t-distribution has $df = n - 1$.

Hypothesis-Testing and Confidence Intervals

We observed that if $\widehat{\theta}$ is an estimator for θ that has an approximately normal sampling distribution, a two-sided confidence interval for θ with confidence coefficient $1-\alpha$ is given by

$$\widehat{\theta} \pm Z_{\alpha/2}\sigma_{\widehat{\theta}}.$$

And note that $P(Z \geq Z_{\alpha/2}) = \alpha/2$.

For large samples two-tailed test at level α, $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$. The test statistic is

$$Z = \frac{\theta - \theta_0}{\sigma_{\widehat{\theta}}}$$

and $H_0 : \theta = \theta_0$ is rejected if and only if $Z \leq -Z_{\alpha/2}$ or $Z \geq Z_{\alpha/2}$. Or equivalently the accept region is

$$RR^c = \{-Z_{\alpha/2} < Z < Z_{\alpha/2}\}.$$

That is, we do not reject $H_0 : \theta = \theta_0$ in favor of the two-tailed alternative if

$$-Z_{\alpha/2} < \frac{\theta - \theta_0}{\sigma_{\widehat{\theta}}} < Z_{\alpha/2}.$$

Restated, the null hypothesis is not rejected (is “accepted”) at level α if

$$\widehat{\theta} - Z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta_0 < \widehat{\theta} - Z_{\alpha/2}\sigma_{\widehat{\theta}}.$$

Hypothesis-Testing and Confidence Intervals

Thus, a duality exists between our large-sample procedures for constructing a $100(1-\alpha)\%$ two-sided confidence interval and for implementing a two-sided hypothesis test with level α:

- Do not reject $H_0 : \theta = \theta_0$ in favor of $H_0 : \theta \neq \theta_0$ if the value θ_0 lies inside a $100(1-\alpha)\%$ confidence interval for θ.
- Reject H_0 if θ_0 lies outside the interval.

Equivalently, a $100(1-\alpha)\%$ two-sided confidence interval can be interpreted as the set of all values of θ_0 for which $H_0 : \theta = \theta_0$ is "acceptable" at level α.

Notice that any value inside the confidence interval is an acceptable value of the parameter. There is not one acceptable value for the parameter but many (indeed, the infinite number of values inside the interval). For this reason, we usually do not *accept* $H_0 : \theta = \theta_0$, even if the value θ_0 falls inside our confidence interval.

Remark. If we use two-sided confidence intervals to test one-tailed hypothesis test at level α, the confidence level should be chosen as $1-2\alpha$.

Hypothesis-Testing and Confidence Intervals

Correspondence between large-sample, one-sided hypothesis tests at level α and one-sided level $1-\alpha$ confidence intervals:

If we desire an α-level test of $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ (an upper-tail test), we should accept the alternative hypothesis if θ_0 is less than a $100(1-\alpha)\%$ lower confidence bound for θ.

Recall. A $1-\alpha$ level one-sided confidence interval of θ is given by

$$[\overline{Y} - t_\alpha \left(\frac{S}{\sqrt{n}}\right), \infty).$$

If the appropriate alternative hypothesis is $H_a : \theta < \theta_0$ (a lower-tail test), you should reject $H_0 : \theta = \theta_0$ in favor of H_a if θ_0 is larger than a $100(1-\alpha)\%$ upper confidence bound for θ.

Recall. A $1-\alpha$ level one-sided confidence interval of θ is given by

$$(-\infty, \overline{Y} + t_\alpha \left(\frac{S}{\sqrt{n}}\right)].$$

p-Value method

Although small values of α are often recommended, the actual value of α to use in an analysis is somewhat arbitrary. Furthermore, software never uses a significance level to conduct hypothesis testing.

Definition. If W is a test statistic, the p-value, or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

Remark.

(1) The p-value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample. It is the probability of observing, just by chance, a test statistic as extreme as or more extreme than the one observed.

(2) The p-value is the smallest value of α for which the null hypothesis can be rejected. Thus, if the p-value $\leq$ the desired value of α, the null hypothesis is rejected for that value of α.

(3) The **smaller** the p-value becomes, the more compelling is the evidence that the **null hypothesis should be rejected**.

p-Value method

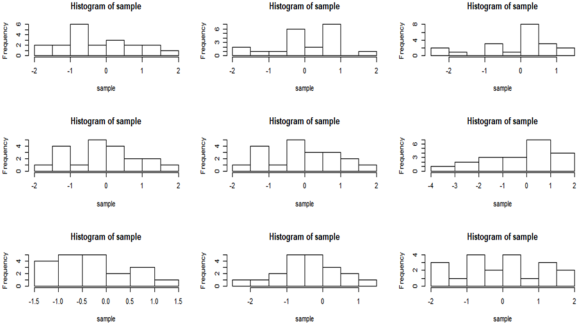
Table 1: Summary of One-Sample Tests at level α: Test statistic, Distribution of the test statistic under H_0, RR and p-value

Population	$N(\mu, \sigma^2)$ or n large	$N(\mu, \sigma^2)$, σ unknown	
Test statistic.	$Z_0 = \frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}}$ or $\frac{\overline{Y}_n - \mu_0}{S/\sqrt{n}}$	$t_0 = \frac{\overline{Y}_n - \mu_0}{S/\sqrt{n}}$	
Distribution.	$Z_0 \sim N(0,1)$	$t_0 \sim t(n-1)$	
One-sided. $H_0 : \mu = \mu_0$ $H_a : \mu > \mu_0$	$RR = \{Z : Z \geq Z_\alpha\}$ p-value=$P(Z \geq Z_0)$	$RR = \{t : t \geq t_\alpha\}$ p-value=$P(t \geq t_0)$	
One-sided. $H_0 : \mu = \mu_0$ $H_a : \mu < \mu_0$	$RR = \{Z : Z \leq -Z_\alpha\}$ p-value=$P(Z \leq Z_0)$	$RR = \{t : t \leq -t_\alpha\}$ p-value=$P(t \leq t_0)$	
Two-sided. $H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	$RR = \{Z : \|Z\| \geq Z_{\alpha/2}\}$ p-value=$2P(Z \geq \|Z_0\|)$	$RR = \{t : \|t\| \geq t_{\alpha/2}\}$ p-value=$2P(t \geq \|t_0\|)$	

Assessing Normality

If the sample size is small, we have to make sure that the data are from a normal population before we construct t-confidence intervals or conduct t-test. Procedure for determining whether it is reasonable to assume that sample data are from a normally distributed population:

- **Histogram**: Construct a histogram. Reject normality if the histogram departs **dramatically** from a bell shape. If the histogram has a small departure from a bell shape, there is insufficient evidence to reject normality. The following shows the histograms of 9 samples of size 20 from the standard normal distribution.



Assessing Normality

- **Outliers**: Identify outliers. Reject normality if there is more than one outlier present.
- **Normal Quantile Plot (QQ plot)**: If the histogram is basically symmetric and there is at most one outlier, use technology to generate a normal quantile plot.
- **Hypothesis test**:If you still cannot decide if the normal quantile plot suggests rejection of normality, you can conduct formal hypothesis tests.

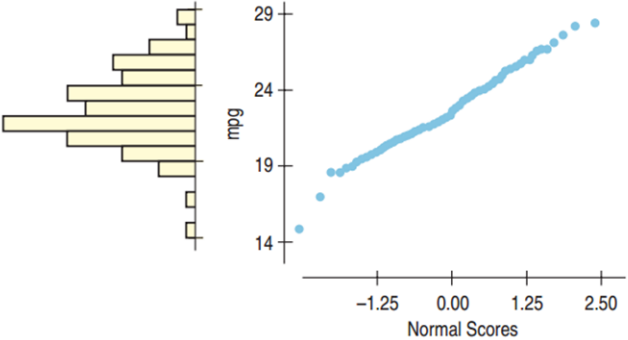
Assessing Normality

Use the following criteria to determine whether or not a data set is from a normal distribution.

Checking Normality Using Normal Quantile Plot

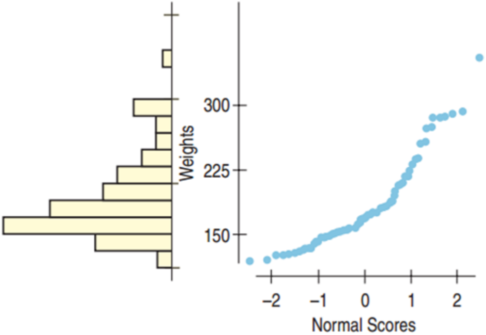
- **Normal Distribution**: The population distribution is normal if the pattern of the points is reasonably close to a straight line and the points do not show some systematic pattern that is not a straight-line pattern.
- **Not a Normal Distribution**: The population distribution is not normal if either or both of these two conditions applies:
  - The points do not lie reasonably close to a straight line.
  - The points show some systematic pattern that is not a straight-line pattern.

Assessing Normality



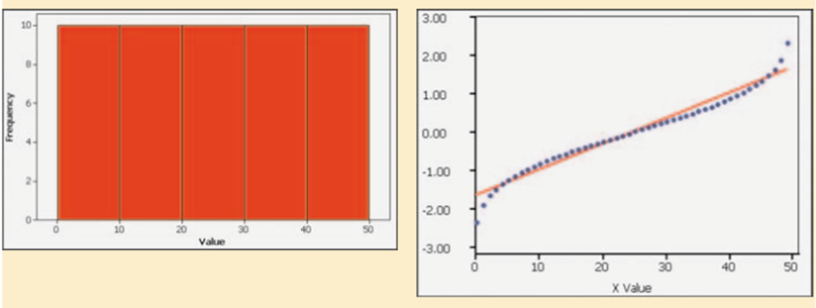
The Normal probability plot is nearly straight, so the Normal model applies. Note that the histogram is unimodal and somewhat symmetric.

Assessing Normality



The Normal probability plot is not straight, so the Normal model does not apply applies. Note that the histogram is skewed to the right.

Assessing Normality



Histogram of data shows a uniform distribution. The corresponding normal quantile plot suggests that the points are not normally distributed because the points show a systematic pattern that is not a straight-line pattern. These sample values are not from a population having a normal distribution.

Assessing Normality

Understanding Q-Q Plots

- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that is roughly straight.
- A normal quantile plot is a graph of points (x, y), where each x value is from the original set of sample data, and each y value is the corresponding z score that is a quantile value expected from the standard normal distribution.

Assessing Normality

Manual Construction of a Normal Quantile Plot:

- **Step 1:** First sort the data by arranging the values in order from lowest to highest.
- **Step 2:** With a sample of size n, each value represents a proportion of $1/n$ of the sample. Using the known sample size n, identify the areas of $1/(n+1), 2/(n+1), 3/(n+1)$ and so on (this is VW method; there are two more methods, BLOM and Tukey, to determine the area to the left of the nth ordered observation). These are the cumulative areas to the left of the corresponding sample values.
- **Step 3:** Use the standard normal distribution to find the z scores corresponding to the cumulative left areas found in Step 2. (These are the z scores that are expected from a normally distributed sample.)
- **Step 4:** Match the original sorted data values with their corresponding z scores found in Step 3, then plot the points (x, y), where each x is an original sample value and y is the corresponding z score.
- **Step 5:** Examine the normal quantile plot and determine whether or not the distribution is normal.

SAS Example

Example. Researchers have shown that cigarette smoking has a deleterious effect on lung function. In their study of the effect of cigarette smoking on the carbon monoxide diffusing capacity (DL) of the lung, Ronald Knudson, W. Kaltenborn and B. Burrows found that current smokers had DL readings significantly lower than either ex-smokers or nonsmokers. The carbon monoxide diffusing capacity for a random sample of current smokers was as follows:

103.768	88.602	73.003	123.086	91.052
92.295	61.675	90.677	84.023	76.014
100.615	88.017	71.210	82.115	89.222
102.754	108.579	73.154	106.755	90.479

Do these data indicate that the mean DL reading for current smokers is lower than 100, the average DL reading for nonsmokers? Test at the $\alpha = 0.01$ level. What is the p-value of the test?

SAS Example

- Normality check

```
data cig;
input dl@@;
cards; /* or use 'datalines;'*/
103.768   88.602   73.003   123.086   91.052
92.295    61.675   90.677   84.023    76.014
100.615   88.017   71.210   82.115    89.222
102.754   108.579   73.154   106.755   90.479
;
run;
```

```
proc univariate data=cig normal plot; /* check normality*/
var dl;
histogram dl/normal;
run;
```

SAS Example

- t-test

```
proc ttest sides=L data=cig H0=100;
var dl;
run;
```

- To get a confidence interval only, we may use `proc means`

```
proc means data=cig alpha=0.02 /*98% confidence level*/
clm mean std;
var dl;
run;
```

License