Chapter 10 Data on a Continuous Variable

We will cover only Sections 10.1.1–10.1.3.

Section 10.1 One-Sample Issues

In Chapters 8 and 9, we learned CIs and significance tests for a population mean μ. Both require that we have a _____ from a _____ population.

If the sample size is large, the assumption of a normal distribution is not so crucial. (Why?)
However, if we do not have much data, this assumption will be important to check. So we have the following question.

How to judge if data follow approximately a normal distribution?

1. **First look at a histogram or stem-leaf plot** - check for non-normal features such as gaps, outliers, and strong skewness.

   If roughly symmetric, unimodal, bell-shaped - then we can turn to a tool that is more sensitive for assessing normality.

2. **Normal Quantile Plot (aka Q-Q Plot or Normal Probability Plot)**

Big Idea: Plot percentiles of a standard normal distribution against the corresponding percentiles of the data.

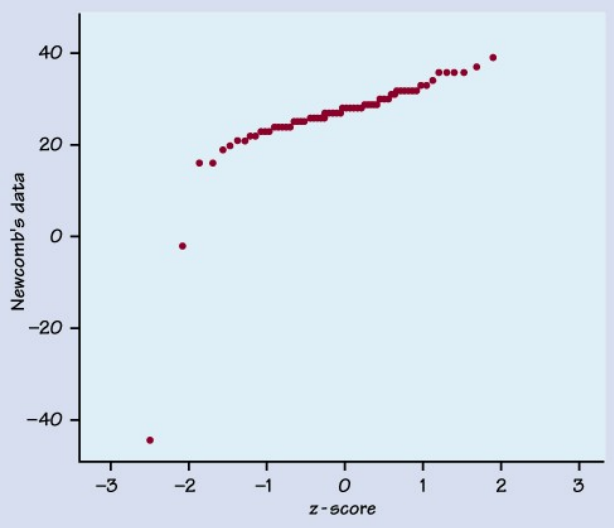
If the observations follow a normal distribution, the resulting plot should be _____

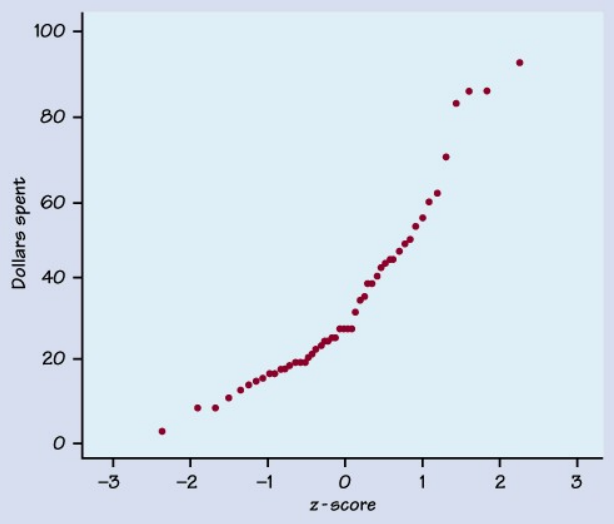
Deviations from this would indicate possible departures from a normal distribution:

- **Outliers** appears as points that are far away from the overall pattern of the plot.
- In a **positively skewed** distribution, the **largest** observations fall distinctly **above** a line drawn through the main body of points.
- In a **negatively skewed** distribution, the _____ observations fall distinctly _____ the line.

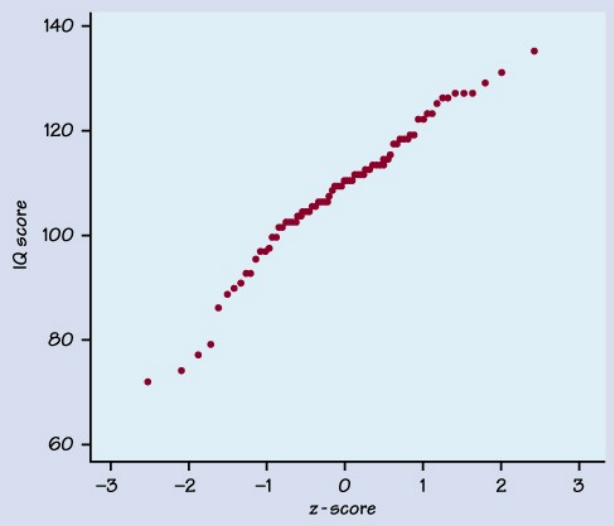
NOTE: Real data almost always show some departure from the "theoretical" normal distribution — don't overreact to minor wiggles in the plot (see Fig. 10.1.3 on page 413).

Examples of Normal Quantile Plots:



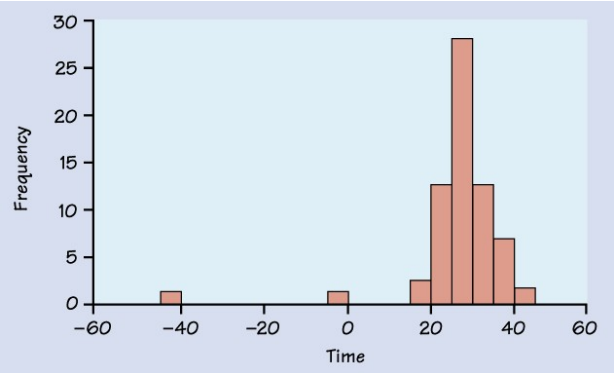


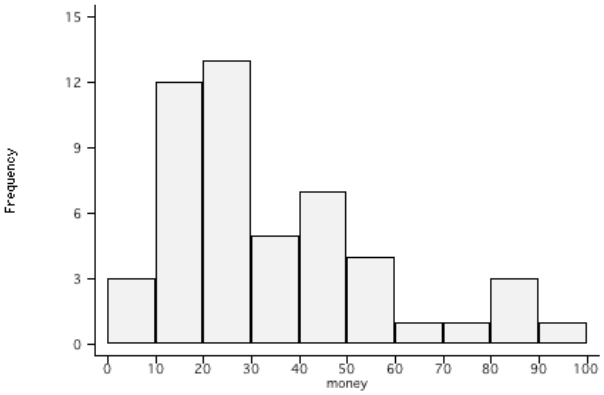
(a) _____ (b) _____



(c) _____

The histograms for (a) and (b) are





Note:

- If the assumptions are false, the results of the analysis may be meaningless.
- The t-tests and CIs are _____ to presence of outliers.
- The two-sided t-tests and CIs are _____ against the departure from the normality assumption if there are no apparent outliers.
- One-sided tests are more sensitive to skewness.
- Do not use t-procedures for small to moderate samples ($n \leq 40$) if there are outliers.
- **Always plot your data before using formal tools of analysis** (tests and confidence intervals).
- Nonparametric methods (to be introduced later) do not assume any particular distribution and are less sensitive to outliers.

Section 10.1.2 Paired Comparisons

Recap: We have discussed the one-sample procedures for estimating a population mean and testing hypotheses about the population mean. One sample designs have a major drawback – **no comparison group** – and thus are subject to various biases (placebo effect, confounding, etc.)

Matched pairs designs are one way to introduce comparison into a study. One matched pairs design has all subjects receive just one treatment, but responses are recorded before and after the treatment. The differences in responses are examined for learning about the effect (if any) of the treatment.

In these paired designs, it is the **differences** that we are interested in analyzing. By focusing on the differences we again have just one sample of observations (the differences) and are able to use the "one-sample" inference methods to form a confidence interval of the mean difference or to test hypotheses about the mean difference.

For paired data, analyze the differences.

Examples: Do piano lessons improve spatial-temporal reasoning of preschool children?

Data: The changes (= after piano lessons - before piano lessons) in reasoning scores for n=34 preschool children:

```
2 5 7 -2 2 7 4 1 0 7 3 4 3 4 9 4 5 2 9 6 0 3 6 -1 3 4 6 7 -2 7 -3 3 4 4
```

The larger values, the better reasoning.

Here is Stata T-Test Output

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   score |      34    3.617647    .5239618    3.055196    2.551639    4.683655
------------------------------------------------------------------------------
Degrees of freedom: 33
```

```
                    Ho: mean(score) = 0

  Ha: mean < 0              Ha: mean ~= 0              Ha: mean > 0
    t =   6.9044              t =   6.9044               t =   6.9044
P < t =   1.0000          P > |t| =   0.0000          P > t =   0.0000
```

(a) Give a 95% CI for the mean improvement in reasoning scores.

(b) State the hypotheses:

(c) What is the observed test statistic?

(d) What are the degrees of freedom?

(e) What is the p-value?

(f) What is the decision?

(g) What is the conclusion?

Section 10.1.3 A Nonparametric Test

We will learn one Nonparametric Test, called **sign test**. It does not assume the data are from normal population, but still assume our data are a random sample.

For nonparametric tests, the hypotheses are about the population **median** $\tilde{\mu}$. The null hypothesis would be $H_0 : \tilde{\mu} = 0$. The alternative can be one-sided or two-sided.

Examples: Do piano lessons improve spatial-temporal reasoning of preschool children?

$H_0 : \tilde{\mu} = 0$.

H_1 :

If H_0 were true, we would expect that the changes are roughly equally likely to be negative and positive. If there are much more positive numbers than negative numbers, we would reject the null. We could compute the P-value indeed.

Counting positive and negative numbers can be done simply by looking at the sign of the changes. So the name of

The Sign Test

For each numbers, use + and − for positive and negative numbers, respectively. For our example, we have

```
+ + + - + + + +  0 + + + + + + + + + + + 0 + + - + + + +  - + - + + +
```

Let X =# of "+". What is the distribution of X if H_0 were true?

For our data, there are 34 observations and two ties (change=0). We can ignore the ties. So X has _____ distribution if H_0 were true.

How many "+" are there? _____

How to find the P-value? Recall that

P-value = the probability of getting a test statistic as extreme or more extreme than the actual observed test statistic, assuming H_0 is true. (The direction of extreme matches H_1.)

How likely will we have as many as _____ or more "+" if H_0 were true? Or

How likely will we have as many as _____ or more heads if a fair coin is tossed 34 times?

The P-value=

Decision:

Conclusion:

Comments on signed tests and nonparametric methods

- less sensitive to outliers
- do not assume any particular distribution for the original observations
- do assume random samples from the populations of interest
- measure of center is the median rather than the mean
- tend to be somewhat less effective at detecting departures from a null hypothesis
- tend to give wider confidence intervals