One-sample normal hypothesis Testing, paired t-test, two-sample normal inference, normal probability plots

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Hypothesis testing

Recall the one-sample normal model

$$Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2).$$

We may perform a t-test to determine whether μ is equal to some specified value μ_0.

The **test statistic** gives information about whether $\mu = \mu_0$ is plausible:

$$t^* = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}.$$

If $\mu = \mu_0$ is true, then $t^* \sim t_{n-1}$.

Hypothesis testing

Rationale: Since $\bar{y}$ is our best estimate of the unknown μ, $\bar{y} - \mu_0$ will be small if $\mu = \mu_0$. But how small is small?

Standardizing the difference $\bar{y} - \mu_0$ by an estimate of $\text{sd}(\bar{Y}) = \sigma/\sqrt{n}$, namely the *standard error* of $\bar{Y}$, $\text{se}(\bar{Y}) = s/\sqrt{n}$ gives us a known distribution for the test statistic t^* *before we collect data*.

$$\boxed{\text{If } \mu = \mu_0 \text{ is true, then } t^* \sim t_{n-1}.}$$

Three types of test

Two sided: $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$.

One sided, "$<$": $H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0$.

One sided, "$>$": $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$.

If the t^* we observe is highly unusual (relative to what we might see for a t_{n-1} distribution), we may reject H_0 and conclude H_a.

Let α be the significance level of the test, the maximum allowable probability of rejecting H_0 when H_0 is true.

Rejection rules

- Two sided: If $|t^*| > t_{n-1}(1 - \alpha/2)$ then reject H_0, otherwise accept H_0.
- One sided, $H_a : \mu < \mu_0$. If $t^* < t_{n-1}(\alpha)$ then reject H_0, otherwise accept H_0.
- One sided, $H_a : \mu > \mu_0$. If $t^* > t_{n-1}(1 - \alpha)$ then reject H_0, otherwise accept H_0.

p-value approach

We can also measure the evidence against H_0 using a p-value, which is the probability of observing a test statistic value *as extreme or more extreme* that the test statistic *we did observe*, if H_0 were true.

A small p-value provides strong evidence against H_0.

Rule: p-value $< \alpha \Rightarrow$ reject H_0, otherwise accept H_0.
p-values are computed according to the alternative hypothesis.
Let $T \sim t_{n-1}$; then

- Two sided: $p = P(|T| > |t^*|)$.
- One sided, $H_a : \mu < \mu_0$: $p = P(T < t^*)$.
- One sided, $H_a : \mu > \mu_0$: $p = P(T > t^*)$.

Example

We wish to test whether the true mean high temperature is greater than 75^o using $\alpha = 0.01$:

$$H_0 : \mu = 75 \text{ versus } H_a : \mu > 75.$$

$$t^* = \frac{77.667 - 75}{8.872/\sqrt{30}} = 1.646 < t_{29}(0.99) = 2.462.$$

What do we conclude?

Note that $p = 0.05525 > 0.01$.

Connection between CI and two-sided test

An α-level two-sided test rejects $H_0 : \mu = \mu_0$ *if and only if* μ_0 falls outside the $(1 - \alpha)100\%$ CI about μ.

Example (continued): Recall the 90% CI for Seattle's high temperature is $(74.91, 80.42)$ degrees

- At $\alpha = 0.10$, would we reject $H_0 : \mu = 73$ and conclude $H_a : \mu \neq 73$?
- At $\alpha = 0.10$, would we reject $H_0 : \mu = 80$ and conclude $H_a : \mu \neq 80$?
- At $\alpha = 0.05$, would we reject $H_0 : \mu = 80$ and conclude $H_a : \mu \neq 80$?

Paired data

When we have two paired samples (when each observation in one sample can be naturally paired with an observation in the other sample), we can use one-sample methods to obtain inference on the **mean difference.**

Example: $n = 7$ pairs of mice were injected with a cancer cell. Mice within each pair came from the same litter and were therefore biologically similar. For each pair, one mouse was given an experimental drug and the other mouse was untreated. After a time, the mice were sacrificed and the tumors weighed.

One-sample inference on differences

Let (Y_{1j}, Y_{2j}) be the pair of control and treatment mice within litter j, $j = 1, \ldots, 7$.

The difference in control versus treatment within each litter is

$$D_j = Y_{1j} - Y_{2j}.$$

If the differences follow a normal distribution, then we have the model

$$D_j = \mu_D + \epsilon_j,\ j = 1, \ldots, n, \quad \text{where } \epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2).$$

Note that μ_D is the mean difference.

To test whether the control results in a higher mean tumor weight, form

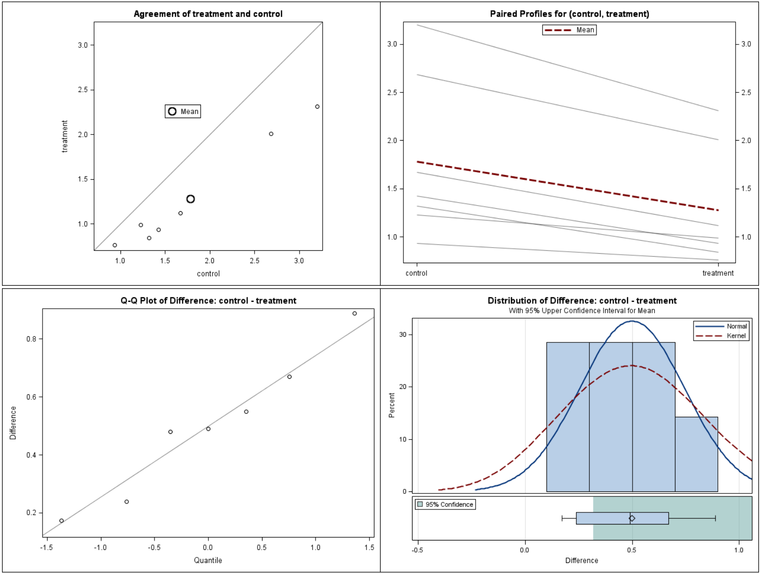
$$H_0 : \mu_D = 0 \text{ versus } H_a : \mu_D > 0.$$

SAS code & output

Actual control and treatment tumor weights are input (not the differences).

```
data mice;
input control treatment @@; * @@ allows us to enter more than two values on each line;
datalines;
1.321 0.841 1.423 0.932 2.682 2.011 0.934 0.762 1.230 0.991 1.670 1.120 3.201 2.312
;
proc ttest h0=0 alpha=0.01 sides=u; * sides=L for Ha: mu_D<0 or sides=2 for two-sided;
paired control*treatment;
run;
                                Difference:  control - treatment

                  N        Mean     Std Dev     Std Err     Minimum     Maximum
                  7      0.4989      0.2447      0.0925      0.1720      0.8890

                     Mean       95% CL Mean        Std Dev      95% CL Std Dev
                   0.4989      0.3191  Infty        0.2447      0.1577   0.5388

                                      DF    t Value    Pr > t
                                       6       5.39    0.0008
```

The default SAS plots...



Agreement of treatment and control

Paired Profiles for (control, treatment)

Q-Q Plot of Difference: control - treatment

Distribution of Difference: control - treatment

With 95% Upper Confidence Interval for Mean

Paired example, continued...

We'll discuss the histogram and the Q-Q plot shortly...

For this test, the p-value is 0.0008. At $\alpha = 0.05$, we reject H_0 and conclude that the true mean difference is larger than 0.

Restated: the treatment produces a significantly lower mean tumor weight.

A 95% CI for the true mean difference μ_D is $(0.27, 0.73)$ (re-run with `sides=2`). The mean tumor weight for untreated mice is between 0.27 and 0.73 grams higher than for treated mice.

Section A.7 Two independent samples

Assume we have two *independent* (not paired) samples from two normal populations. Label them 1 and 2. The model is

$$Y_{ij} = \mu_i + \epsilon_{ij}, \text{ where } i = 1, 2 \text{ and } j = 1, \ldots, n_i.$$

The “within sample heterogeneity” follows

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

- Both populations have the same variance σ^2.
- The two sample sizes (n_1 and n_2) may be different.

Pooled approach

An estimator of the variance σ^2 is the “pooled sample variance”

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Then

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim t_{n_1+n_2-2}.$$

We are interested in the mean difference $\mu_1 - \mu_2$, i.e. the difference in the population means.

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{n_1+n_2-2}(1-\alpha/2)\sqrt{s_p^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}.$$

Pooled approach: Hypothesis test

Often we wish to test whether the two populations have the same mean, i.e. $H_0 : \mu_1 = \mu_2$. Of course, this implies $H_0 : \mu_1 - \mu_2 = 0$. The test statistic is

$$t^* = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_p^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}},$$

and is distributed $t_{n_1+n_2-2}$ under H_0. Let $T \sim t_{n_1+n_2-2}$. The tests are carried out via:

H_a	Rejection rule	p-value
$\mu_1 \neq \mu_2$	$\lvert t^*\rvert > t_{n_1+n_2-2}(1-\alpha/2)$	$P(\lvert T\rvert > \lvert t^*\rvert)$
$\mu_1 < \mu_2$	$t^* < -t_{n_1+n_2-2}(1-\alpha)$	$P(T < t^*)$
$\mu_1 > \mu_2$	$t^* > t_{n_1+n_2-2}(1-\alpha)$	$P(T > t^*)$

Unequal variances: Satterthwaite approximation

Often, the two populations have quite different variances. Suppose $\sigma_1^2 \neq \sigma_2^2$. The model is

$$Y_{ij} = \mu_i + \epsilon_{ij}, \ \ \epsilon_{ij} \overset{ind.}{\sim} N(0, \sigma_i^2)$$

where $i = 1, 2$ denotes the population and $j = 1, \ldots, n_i$ the measurement within the population.

Use s_1^2 and s_2^2 to estimate σ_1^2 and σ_2^2. Define the test statistic

$$t^* = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Unequal variances: Satterthwaite approximation

Under the null $H_0 : \mu_1 = \mu_2$, this test statistic is *approximately* distributed $t^* \sim t_{df}$ where

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Note that $df = n_1 + n_2 - 2$ when $n_1 = n_2$ and $s_1 = s_2$.

Satterthwaite and pooled variance methods typically give similar results when $s_1 \approx s_2$.

Testing $H_0 : \sigma_1 = \sigma_2$

- We can formally test $H_0 : \sigma_1 = \sigma_2$ using various methods (e.g. Bartlett's F-test or Levene's test), but in practice graphical methods such as box plots are often used.
- SAS automatically provides the "folded F test"

$$F^* = \frac{\max\{s_1^2, s_2^2\}}{\min\{s_1^2, s_2^2\}}.$$

- This test assumes normal data and is sensitive to this assumption.

SAS example

Example: Data were collected on pollution around a chemical plant (Rao, p. 137). Two independent samples of river water were taken, one upstream and one downstream. Pollution level was measured in ppm. Do the mean pollution levels differ at $\alpha = 0.05$?

SAS code

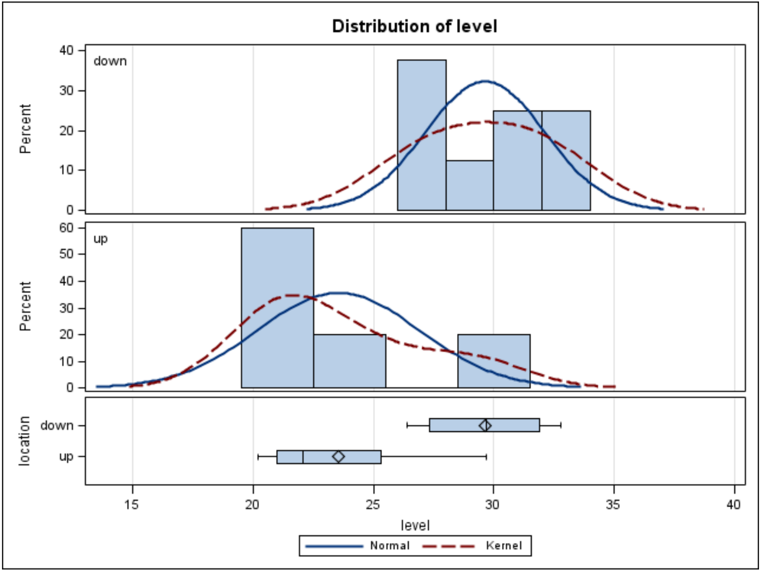
```
data pollution;
input level location $ @@; * use $ to denote a categorical variable;
datalines;
24.5 up   29.7 up   20.4 up   28.5 up   25.3 up   21.8 up   20.2   up 21.0 up
21.9 up   22.2 up   32.8 down 30.4 down 32.3 down 26.4 down 27.8 down 26.9 down
29.0 down 31.5 down 31.2 down 26.7 down 25.6 down 25.1 down 32.8 down 34.3 down
35.4 down
;

proc ttest h0=0 alpha=0.05 sides=2;
 class location; var level;
run;
```

SAS output

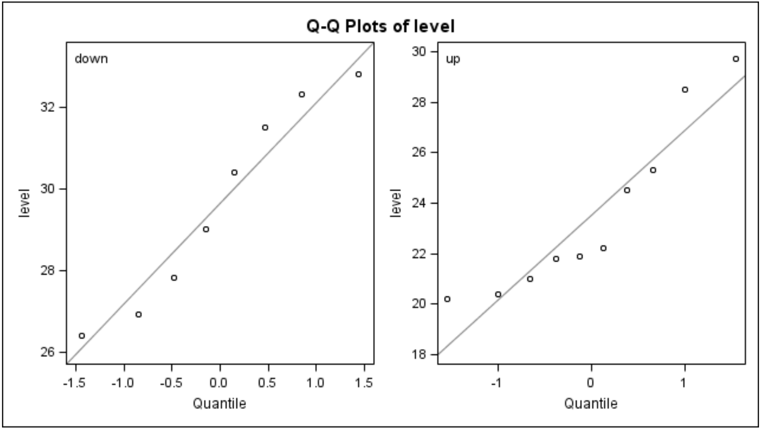
```
                               The TTEST Procedure

                                 Variable:  level

location        N        Mean     Std Dev     Std Err     Minimum     Maximum
down            8     29.6375      2.4756      0.8752     26.4000     32.8000
up             10     23.5500      3.3590      1.0622     20.2000     29.7000
Diff (1-2)             6.0875      3.0046      1.4252

location      Method              Mean      95% CL Mean      Std Dev    95% CL Std Dev
down                           29.6375    27.5679  31.7071    2.4756    1.6368   5.0384
up                             23.5500    21.1471  25.9529    3.3590    2.3104   6.1322
Diff (1-2)    Pooled            6.0875     3.0662   9.1088    3.0046    2.2377   4.5728
Diff (1-2)    Satterthwaite     6.0875     3.1687   9.0063

Method           Variances        DF    t Value    Pr > |t|
Pooled           Equal            16       4.27      0.0006
Satterthwaite    Unequal      15.929       4.42      0.0004

              Equality of Variances

Method      Num DF    Den DF    F Value    Pr > F
Folded F         9         7       1.84    0.4332
```

Note that $1.84 = 3.3590^2/2.4756^2$.

The default SAS plots...



The default SAS plots...



Assumptions going into t procedures

Note: Recall our t-procedures require that the data come from normal population(s).

Fortunately, the t procedures are **robust**: they work approximately correctly if the population distribution is “close” to normal.

Also, if our sample sizes are large, we can use the t procedures (or simply normal-based procedures) even if our data are not normal because of the central limit theorem.

If the sample size is small, we should perform some check of normality to ensure t tests and CIs are okay.

Question: Are there any other assumptions going into the model that can or should be checked or at least thought about? For example, what if pollution measurements were taken on consecutive days?

Boxplots for checking normality

- To use t tests and CIs in small samples, approximate normality should be checked.
- Could check with a histogram or boxplot: verify distribution is *approximately* symmetric.
- **Note**: For perfectly normal data, the probability of seeing an outlier on an R or SAS boxplot using defaults is 0.0070. For a sample size $n_i = 150$, in perfectly normal data we expect to see $0.007(150) \approx 1$ outlier. Certainly for small sample sizes, we expect to see none.

Q-Q plot for checking normality

- More precise plot: normal Q-Q plot. Idea: the human eye is very good at detecting deviations from linearity.
- Q-Q plot is ordered data against normal quantiles $z_i = \Phi^{-1}\{i/(n+1)\}$ for $i = 1, \dots, n$.
- Idea: $z_i \approx E(Z_{(i)})$, the expected order statistic under standard normal assumption.
- A plot of $y_{(i)}$ versus z_i should be reasonably straight *if data are normal*.
- However, in small sample sizes there is a lot of variability in the plots even with perfectly normal data...

Example: R example of Q-Q plots and R code to examine normal, skewed, heavy-tailed, and light-tailed distributions. Note both Q-Q plots and numbers of outliers on R boxplot.

Assumptions for two examples

- Mice tumor data: Q-Q plot? Boxplot? Histogram?
- River water pollution data: Q-Q plots? Boxplots? Histograms?

Teaching effectiveness, continued...

```
proc ttest h0=0 alpha=0.05 sides=2;
 class attend; var rating;
run;
```