Some Statistics Background Needed for STA 302/1001

A.L. Gibbs

1 A Brief Review of the Distribution Theory for t-tests and Confidence Intervals

Some facts related to Normally distributed random variables:

1. Consider a random variable X whose distribution is $N(\mu, \sigma^2)$. To standardize X, let

$$Z = \frac{X - \mu}{\sigma}$$

then $Z \sim N(0, 1)$.

2. Any linear combination of Normally distributed random variables is also normally distributed.

3. If U and V are independent random variables with $U \sim N(0, 1)$ and $V \sim \text{chisquare}(m)$ then $\frac{U}{\sqrt{V/m}}$ has a t distribution with m degrees of freedom.

4. If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables then

   (a) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an (unbiased) estimator of μ.

   (b) An (unbiased) estimator of σ^2 is

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

   and $(n-1)S^2/\sigma^2$ has a chisquare distribution with $n-1$ degrees of freedom.

   (c) S (the square root of S^2) and $\overline{X}$ are independent.

   (d)

   $$\frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{s/\sigma} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

2 Confidence Intervals for the Mean of a Normal Distribution

Suppose $x_1, x_2, \ldots, x_n$ are realizations of i.i.d. random variables $X_1, X_2, \ldots, X_n$ which have the $N(\mu, \sigma^2)$ distribution.

Then $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ and

$$\Pr\left(\left|\frac{\overline{X}-\mu}{S/\sqrt{n}}\right| \leq t_{n-1,\,\alpha/2}\right) = 1-\alpha$$

where $t_{n-1,\,\alpha/2}$ is the value from the t_{n-1} distribution such that $\alpha/2$ is the probability above it, i.e. it is the $1-\alpha/2$ quantile from the t_{n-1} distribution.

The interval

$$\overline{x} \pm t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ confidence interval for μ where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

Note that the form of the C.I. is *estimate* $\pm$ *quantile* $\times$ *standard error*.

How to interpret the C.I.:

> Confidence intervals calculated by this method from repeated samples from a $N(\mu, \sigma^2)$ distribution of size n will include the true (unknown) value of μ $100(1-\alpha)\%$ of the time.

A common misinterpretation of C.I.s:

> The probability that μ is in the interval is $100(1-\alpha)\%$.

What is the error in this misinterpretation?

Confidence intervals are given to give an idea of the precision of an estimate of a parameter.

Usual values of α:
0.01 (gives 99% C.I.)
0.05 (gives 95% C.I.)
0.10 (gives 90% C.I.)

3 Steps for Hypothesis Testing

Note that the focus here is on getting and interpreting p-values and **not** *on rejection regions, which are useful for theoretical analysis but are limiting in practice.*

1. Establish null (H_0) and alternative (H_a) hypotheses for the value of the parameter of interest. Typically the alternative hypothesis is what is of interest.

2. Calculate a test statistic whose distribution is known assuming that the null hypothesis is true.

3. Estimate the p-value. Assuming the null hypothesis is true, the p-value is the probability of the value of the test statistic that was observed or a value more extreme (where "more extreme" values belong to the alternative hypothesis). The p-value is a measure of the strength of the evidence against H_0 in favour of H_a.

4. If the p-value is small, then either:
   (1) H_0 is correct and the observed data happened to be one of those rare samples that produces an unusual test statistic (Type I error)
   or
   (2) H_0 is incorrect.
   The smaller the p-value the stronger the evidence that H_0 is incorrect. A large p-value indicates that the data are consistent with H_0 (which doesn't necessarily mean that H_0 is true).

How small is "small"? The boundaries are grey, but here are some typical guidelines:

$p > 0.1$	No evidence against H_0
$0.05 < p < 0.1$	Some weak evidence against H_0 (suggestive but inconclusive)
$0.01 < p < 0.05$	Moderate evidence against H_0
$p < 0.01$	Strong evidence against H_0

4 Tests for Comparing the Means of Two Normal Distributions

Probably the most commonly carried out tests in statistics are tests to compare whether two independent samples are from distributions with the same mean, assuming the distributions are normal. Even if they aren't normal distributions, the tests are very robust since all sample means are approximately normally distributed by the Central Limit Theorem.

Suppose we have a sample of size n_1 from a random variable $X_1 \sim N(\mu_1, \sigma_1^2)$ (i.e. n_1 independent realizations of X_1) and a sample of size n_2 from a random variable $X_2 \sim N(\mu_2, \sigma_2^2)$. Comparing whether $\mu_1 = \mu_2$ is equivalent to testing if $\mu_1 - \mu_2 = 0$ so we are interested in estimating $\mu_1 - \mu_2$ for which we'll use the (unbiased) estimator $\overline{X}_1 - \overline{X}_2$. The distribution of the estimator is

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

To test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$, the two independent sample t-test has as test statistic $(\overline{x}_1 - \overline{x}_2)$/the standard error of $(\overline{x}_1 - \overline{x}_2)$. There are two common approaches to calculating the standard error.

1. *Assume $\sigma_1 = \sigma_2$.*
   Then the test statistic
   $$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$
   where s_p is the pooled standard deviation
   $$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

2. *Don't assume the standard deviations are equal.*
   Then the test statistic
   $$t_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
   has approximately a t-distribution with the degrees of freedom estimated by the Satterthwaite approximation
   $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$