Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 12: t tests for two groups

Lecturer: Dominik Rothenhäusler *February 27*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

12.1 Recap

This lecture was about t-tests to compare two means. We began by wrapping up where the t test comes from. That is described in the Lecture 10 notes. We ended up with

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

when $X_i \sim N(\mu, \sigma^2)$ for $n \geqslant 2$ and $\sigma > 0$.

For $0 < \gamma < 1$ and $\nu > 0$, let $t_{(\nu)}^{\gamma}$ be the γ quantile of the $t_{(\nu)}$ distribution. That is

$$\Pr(t_{(\nu)} \leqslant t_{(\nu)}^{\gamma}) = \gamma. \tag{12.1}$$

Here ν represents the degrees of freedom. It can be n or $n-1$ or $n-p$ or $n+m-2$ in different places where it appears. We give it the Greek letter ν ('nu') because nu is like n and the degrees of freedom are typically an adjusted version of n. Now, working in slow motion and using symmetry of the t distribution,

$$\begin{aligned}
\Pr(|t_{(\nu)}| \leqslant t_{(\nu)}^{1-\alpha/2}) &= 1 - \Pr(t_{(\nu)} < -t_{(\nu)}^{1-\alpha/2}) + \Pr(t_{(\nu)} > t_{(\nu)}^{1-\alpha/2}) \\
&= 1 - \bigl(1 - \Pr(t_{(\nu)} > t_{(\nu)}^{1-\alpha/2})\bigr) + \Pr(t_{(\nu)} > t_{(\nu)}^{1-\alpha/2}) \\
&= 2\Pr(t_{(\nu)} > t_{(\nu)}^{1-\alpha/2}) \\
&= 2\bigl(1 - \Pr(t_{(\nu)} \leqslant t_{(\nu)}^{1-\alpha/2})\bigr) \\
&= 2\bigl(1 - (1-\alpha/2)\bigr) \\
&= 2\bigl(\alpha/2)\bigr) \\
&= \alpha.
\end{aligned}$$

What just happened? First we got rid of $-t_{(\nu)}^{1-\alpha/2}$ so that only $+t_{(\nu)}^{1-\alpha/2}$ is involved. Then we have to get it in terms of "$t_{(\nu)} \leqslant$ something" in order to use the defining property (12.1) of quantiles. Maybe it would have been faster to just show that $\Pr(|t_{(\nu)}| > t_{(\nu)}^{1-\alpha/2}) = \alpha$. (Try it.) Keep in mind that α is a small number like 0.01 or 5×10^{-8} or sometimes 0.05 (which is losing its status as the default small probability).

Now for IID $N(\mu, \sigma^2)$ data we have a $100(1-\alpha)\%$ confidence interval for μ of the form

$$\bar{X} \pm \frac{st^{1-\alpha/2}_{(n-1)}}{\sqrt{n}},$$

because

$$\Pr\Bigl(\bar{X} - \frac{st^{1-\alpha/2}_{(n-1)}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{st^{1-\alpha/2}_{(n-1)}}{\sqrt{n}}\Bigr) = 1-\alpha.$$

Notice that $\nu = n-1$ here.

In class I sketched a bell shaped t distribution with area $\alpha/2$ to the right of $t^{1-\alpha/2}_{(\nu)}$, area $\alpha/2$ to the left of $-t^{1-\alpha/2}_{(\nu)}$ and area $1-\alpha$ in between. For a symmetric distribution like the t distribution we know that $\Pr(t > x) = \Pr(t < -x)$.

We can also write our confidence interval as

$$\Pr\Bigl(\bar{X} + \frac{st^{\alpha/2}_{(n-1)}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{st^{1-\alpha/2}_{(n-1)}}{\sqrt{n}}\Bigr) = 1-\alpha.$$

This works because $t^{\alpha/2}_{(n-1)} = -t^{1-\alpha/2}_{(n-1)}$. This version makes it clear that we are lopping off $\alpha/2$ probability from each tail.

12.2 Why is it $n-1$ in the denominator of s^2?

Let $X_1, \dots, X_n$ be IID with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 > 0$ and $n \geqslant 2$. We customarily estimate σ^2 by

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

instead of

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

If n is large then there is no practical difference between these choices. If n is small they're quite different. For any $n \geqslant 2$ there is a theoretical simplification to using s^2. It is a big improvement for small n and we might as well keep it for large n (where it doesn't matter) in order to avoid picking an awkward switching point between 'large' and 'small' n.

One reason to use s^2 is that

$$\mathbb{E}(s^2) = \sigma^2$$

(so it is unbiased) while $\mathbb{E}(\hat{\sigma}^2) = (1-1/n)\sigma^2$, which is then biased low. To see why it is biased low notice that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu)\sum_{i=1}^n (X_i - \bar{X}) + (\bar{X} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 + (\bar{X} - \mu)^2,
\end{aligned}$$

because X_i sum to $n\bar{X}$. We use this subtract and add technique repeatedly in statistics. The left hand side has expected value σ^2 and $\mathbb{E}((\bar{X}-\mu)^2) = \sigma^2/n$. Subtracting yields

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2 - \frac{1}{n}\sigma^2 = \Bigl(1 - \frac{1}{n}\Bigr)\sigma^2$$

and now we see that dividing by $n-1$ instead of n fixes the bias giving $\mathbb{E}(s^2) = \sigma^2$.

If $n = 1$ we are out of luck as far as estimating σ^2 goes.

Another way to see this is that if we use calculus to minimize

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

over m the answer is $m = \bar{X}$. If we knew μ we could use

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

but plugging $\bar{X}$ in for μ gives us the smallest possible sum of squares and it is biased low.

12.3 Exploring the t table

We can explore the t quantiles. This function:

```
ttables = function( df = c(1,3,5,10,100,1000,1/0),alpha=c(.05,.01,.001,.0001,.00001,1e-8)){

n = length(df)
p = length(alpha)

ans = matrix(0,n,p)
for( i in 1:n )
for( j in 1:p )
  ans[i,j] = qt( 1-alpha[j]/2,df[i] )

colnames(ans)=alpha
rownames(ans)=df
ans[n:1,]
}
```

gives this output

```
> source("tables.R");signif( ttables(), 3 )
      0.05  0.01  0.001   1e-04    1e-05    1e-08
Inf   1.96  2.58   3.29    3.89     4.42 5.73e+00
1000  1.96  2.58   3.30    3.91     4.44 5.78e+00
100   1.98  2.63   3.39    4.05     4.65 6.25e+00
10    2.23  3.17   4.59    6.21     8.15 1.71e+01
5     2.57  4.03   6.87   11.20    17.90 7.17e+01
3     3.18  5.84  12.90   28.00    60.40 6.04e+02
1    12.70 63.70 637.00 6370.00 63700.00 6.37e+07
```

Let's first look at the upper left corner: for normally distributed data $\pm$ two standard errors gets you about 95% confidence. The top row shows that even a small increase in the number of standard errors pushes the confidence level very close to 100%. The normal distribution has light tails. The first column shows that as the number ν of degrees of freedom drops you have to use only slightly more than $\pm$ two standard errors until the df gets low. Then it grows. The bottom row is a mild horror story about $t_{(1)}$, the Cauchy distribution, which has heavy tails.

12.4 Two sample (unpaired) t-test

Suppose that we have $X_1, \dots, X_n \sim F$ (IID) independent of $Y_1, \dots, Y_m \sim G$ (IID) and we wonder whether $\mathbb{E}(X) = \mathbb{E}(Y)$. In class the example was weight loss where X was for diet and Y was for diet plus exercise. We think Y should have a larger mean. But it might not be large enough to matter. Or it could be that after exercise some people don't stick as well to their diet plans.

The simplest case is a model where $X_i \sim N(\mu_x, \sigma^2)$ and $Y_j \sim N(\mu_y, \sigma^2)$. Notice that both groups are assumed to have the same σ. Rice describes what to do otherwise.

Rice works out the generalized likelihood ratio test for this case and winds up with a t test. We will derive the exact same t test motivated by intuition and backed up by distribution theory. You should be comfortable with both approaches.

We could reasonably estimate $\mu_x - \mu_y$ by $\bar{X} - \bar{Y}$. We anticipate a negative value for the running example. Maybe it could come out negative by chance when $\mu_x - \mu_y = 0$. We know that

$$\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \sigma^2(1/n + 1/m)).$$

If only we knew σ we could form a confidence interval using

$$\Pr(|(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)| > Z^{1-\alpha/2}\sigma\sqrt{1/m + 1/n}) = 1 - \alpha,$$

where $Z^{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of the $N(0, 1)$ distribution. This uses symmetry of $N(0, 1)$ just like we used symmetry for t.

Not knowing σ we decide to estimate it. Let

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \text{and} \quad s_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

From normal theory we know that $(n-1)s_x^2/\sigma^2 \sim \chi^2_{(n-1)}$ independently of $(m-1)s_y^2/\sigma^2 \sim \chi^2_{(m-1)}$. Therefore

$$(n-1)s_x^2/\sigma^2 \sim \chi^2_{(n-1)} + (m-1)s_y^2/\sigma^2 \sim \chi^2_{(m-1)} \sim \chi^2_{(n+m-2)}.$$

We can estimate σ^2 by pooling these two estimates into

$$s^2_{\text{pooled}} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$

Now consider

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{s_{\text{pooled}}\sqrt{1/n + 1/m}}.$$

This is the difference between the observed and true difference divided by an estimate of the standard deviation of that difference. Now write it as

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sigma\sqrt{1/n + 1/m}} \Bigg/ \frac{s_{\text{pooled}}}{\sigma}.$$

The numerator is $N(0,1)$. The denominator is $\sqrt{\chi^2_{n+m-2}/(n+m-2)}$. All four of $\bar{X}$, $\bar{Y}$, s_x and s_y are independent, so the denominator is independent of the numerator. As a result t above has the $t_{(n+m-2)}$ distribution.

We reject $H_0 : \mu_x - \mu_y$ in favor of $H_A : \mu_x \neq \mu_y$ at the level α if $|t| > t^{1-\alpha/2}_{(n+m-2)}$. Rice shows that the GLRT also rejects H_0 for large values of $|t|$.

Suppose that we care about $\Delta = \mu_x - \mu_y$ and estimate it by $\hat{\Delta} = \bar{X} - \bar{Y}$. Our null is $H_0 : \Delta = 0$ and the alternative is $H_A : \Delta \neq 0$. Our confidence interval for Δ is

$$\hat{\Delta} \pm t^{1-\alpha/2} s_{\text{pooled}} \sqrt{1/n + 1/m}.$$

Notice that for a narrow confidence interval we really need both n ***and*** m to be large.

12.5 Unequal variances

If $\sigma_x^2 \neq \sigma_y^2$ then things are more complicated. The variance of $\bar{X} - \bar{Y}$ is $\sigma_x^2/n + \sigma_y^2/m$ and

$$\mathbb{E}\Big(s^2_{\text{pooled}}\big(\frac{1}{n} + \frac{1}{m}\big)\Big) = \frac{(n-1)\sigma_x^2 + (m-1)\sigma_y^2}{n+m-2}\Big(\frac{1}{n} + \frac{1}{m}\Big).$$

The actual variance of $\bar{X} - \bar{Y}$ is a weighed sum of σ_x^2 and σ_y^2 with weights inversely proportional to sample sizes n and m. The expected value of the pooled estimate of that variance is also a weighed sum but the weights are proportional to $n-1$ and $m-1$ almost exactly the opposite of what they should be. The t test will not be correct if $\sigma_x \neq \sigma_y$.

What we should do instead is use

$$t' = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{12.2}$$

Squaring the denominator of t' gives an unbiased estimate of $\mathrm{Var}(\bar{X} - \bar{Y})$. It's called t' above instead of t because it does not have a t distribution. It has approximately the t distribution. The degrees of freedom in the approximation are somewhere between $\min(n-1, m-1)$ and $n+m-2$. See ch 11.2 of Rice and Example C there.

12.6 Power

The power is the probability of rejecting the null under the alternative. Rice ch 11.2.2 goes back to the known σ setting to show how power works. Define $\Delta = \mu_x - \mu_y$. Suppose we want to test $H_0 : \Delta = 0$ versus $H_0 : \Delta \neq 0$. For a two sided test with known σ the power when Δ is the true difference is (working in slow motion)

$$
\begin{aligned}
&\Pr(|\bar{X}-\bar{Y}| > Z^{1-\alpha/2}\sigma\sqrt{1/n+1/m})\\
&=\Pr\big(|N(\Delta,\sigma^2/(1/n+1/m))| > Z^{1-\alpha/2}\sigma\sqrt{1/n+1/m}\big)\\
&=\Pr\Big(\Big|N\Big(\frac{\Delta}{\sigma\sqrt{1/n+1/m}},1\Big)\Big| > Z^{1-\alpha/2}\Big)\\
&=\Pr\Big(N\Big(\frac{\Delta}{\sigma\sqrt{1/n+1/m}},1\Big) > Z^{1-\alpha/2}\Big)\\
&\quad+\Pr\Big(N\Big(\frac{\Delta}{\sigma\sqrt{1/n+1/m}},1\Big) < -Z^{1-\alpha/2}\Big)\\
&=\Pr\Big(N(0,1) > Z^{1-\alpha/2}-\frac{\Delta}{\sigma\sqrt{1/n+1/m}}\Big)\\
&\quad+\Pr\Big(N(0,1) < -\frac{\Delta}{\sigma\sqrt{1/n+1/m}}-Z^{1-\alpha/2}\Big)\\
&=\Pr\Big(N(0,1) < -Z^{1-\alpha/2}+\frac{\Delta}{\sigma\sqrt{1/n+1/m}}\Big)\\
&\quad+\Pr\Big(N(0,1) < -\frac{\Delta}{\sigma\sqrt{1/n+1/m}}-Z^{1-\alpha/2}\Big)\\
&=\Phi\Big(-Z^{1-\alpha/2}+\frac{\Delta}{\sigma\sqrt{1/n+1/m}}\Big)+\Phi\Big(-Z^{1-\alpha/2}-\frac{\Delta}{\sigma\sqrt{1/n+1/m}}\Big).
\end{aligned}
$$

If $\Delta = 0$ we get α again as we should. Notice that Δ and σ only enter through the ratio Δ/σ. That is the number of standard deviations separating μ_x and μ_y. If we replaced all our X and Y values by $10X$ and $10Y$ respectively, both Δ and σ would go up by 10 and the effect would cancel.

Suppose we had $\alpha = 0.01$. Then $Z^{1-\alpha/2} = 2.58$. Now our power is

$$\Phi\Big(-2.58+\frac{\Delta}{\sigma\sqrt{1/n+1/m}}\Big)+\Phi\Big(-2.58-\frac{\Delta}{\sigma\sqrt{1/n+1/m}}\Big).$$

If $\Delta > 0$ and n, m both grow to infinity then the first term approaches 1. If $\Delta < 0$ and n, m both grow to infinity then the second term approaches 1. If only one grows then the power does not go to 1. Suppose we are designing a study. We may choose to use $n = m$. Then the power is

$$\Phi\Big(-2.58+\sqrt{\frac{n}{2}}\frac{\Delta}{\sigma}\Big)+\Phi\Big(-2.58-\sqrt{\frac{n}{2}}\frac{\Delta}{\sigma}\Big).$$

In class I mentioned the ***noncentral*** t distribution. That is the key to getting the exact power for the unknown σ case. Is is also what the better online power calculators use. Thinking it over some more, I believe that we are better off not covering it in this class. We have enough ideas in the mix already and the known σ case gets us the insights we need. The differences are minor when there are enough degrees of freedom in the data.

12.7 Paired comparisons

The previous comparison was for two independent samples. In class we considered 'grip strength', a measure of how much pressure a person can generate with one hand. We could expect it to be high for rock climbers but low for frail people. Suppose that we want to compare left versus right handed grip strength. We anticipate enormous differences from person to person and much smaller differences between left and right hands for a given person.

It is obviously more sensible to use left and right hands on the same set of people. (More about that later.) Then a small persistent left vs right hand difference could still be detected. We would get data (X_i, Y_i) for $i = 1, \dots, n$ where X_i was from the left hand and Y_i was from the right hand. We cannot use the previous two sample test because it is very unreasonable to think that X_i is independent of Y_i. We could reasonably assume that (X_i, Y_i) is independent of (X_j, Y_j) for $i \neq j$.

Suppose that (X_i, Y_i) are sampled from a bivariate normal distribution. We write

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right)$$

where $X \sim N(\mu_x, \sigma_x^2)$, and $Y \sim N(\mu_y, \sigma_y^2)$ and they are bivariate normal with correlation ρ. The paired test works with

$$D_i = X_i - Y_i \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y).$$

Let's write this as

$$D_i \sim N(\mu_D, \sigma_D^2).$$

Now we get hypothesis tests and confidence intervals about $\mu_D = \mu_x - \mu_y$ using

$$t = \frac{\bar{D} - \mu_d}{S_{\bar{D}}} = \frac{\bar{D} - \mu_d}{S_D/\sqrt{n}} \sim t_{(n-1)}$$

where $s_D^2 = \frac{1}{n-1}\sum_{i=1}^n (D_i - \bar{D})^2$ (and $s_{\bar{D}}^2 = s_D^2/n$).

It is an enormous conceptual error to make the wrong choice between a paired and unpaired comparison. In the paired setting each X_i is logically related to one Y_i and vice versa (here left and right hands of the same people). In the unpaired setting there is no such logical connection and the two groups of data are unrelated.

If we sampled two different people we would have

$$X_i - Y_j \sim N(\mu_x - \mu_u, \sigma_x^2 + \sigma_y^2).$$

Sampling from the same person

$$X_i - Y_i \sim N(\mu_x - \mu_u, \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y).$$

We anticipate a large positive ρ for grip strength, so the paired analysis has much smaller variance than the unpaired one.

Students in class made two key observations. First, we didn't have to assume that $\sigma_x = \sigma_y$ here. Second, the degrees of freedom are $n - 1$ instead of $n + m - 2$, so we get fewer degrees of freedom.

Choosing an unpaired analysis of data that should have a paired analysis or vice versa is a major conceptual error.