PUBH 7405 - Block 2

LAND ACKNOWLEDGEMENT

The School of Public Health at the University of Minnesota Twin Cities is built within the traditional homelands of the Dakota people. Minnesota comes from the Dakota name for this region, Mni Sóta Maḳoce, which loosely translates to the land where the waters reflect the skies.

It is important to acknowledge the peoples on whose land we live, learn, and work as we seek to improve and strengthen our relations with our tribal nations. We also acknowledge that words are not enough. We must ensure that our institution provides support, resources, and programs that increase access to all aspects of higher education for our American Indian students, staff, faculty, and community members.

Tests Based on Population Models

Getting null distribution based on randomization can be difficult if:

- experiment is complicated
- experiment is non or partially randomized
- experiment includes nuisance factors

Consider the following model for our 2 treatment experiment:

- There is a large/infinite population of similar individuals as those in our experiment.
- When treatment A is given, the distribution of outcomes can be represented by probability distribution p_A:

$$\mathbb{E}(Y_A) = \int y p_A(y)\, dy = \mu_A$$

$$\text{var}(Y_A) = \mathbb{E}[(Y_A - \mu_A)^2] = \sigma_A^2$$

- When treatment B is given, we have probability distribution p_B:

$$\mathbb{E}(Y_B) = \mu_B$$

$$\text{var}(Y_B) = \sigma_B^2$$

- The individuals that got treatment A in our experiment can be viewed as an independent sample from P_A.
- Similarly for those receiving treatment B.

$$Y_{1A}, \ldots, Y_{n_A A} \sim P_A$$

$$Y_{1B}, \ldots, Y_{n_B B} \sim P_B$$

Recall

$$\mathbb{E}\left(\bar{Y}_A\right) = \mathbb{E}\left(\frac{1}{n_A}\sum_{i=1}^{n_A} Y_{iA}\right) = \frac{1}{n_A}\sum_{i=1}^{n_A}\mu_A = \mu_A$$

$$\text{var}\left(\bar{Y}_A\right) = \text{var}\left(\frac{1}{n_A}\sum_{i=1}^{n_A} Y_{iA}\right) = \frac{1}{n_A^2}\sum_{i=1}^{n_A}\sigma_A^2 = \frac{\sigma_A^2}{n_A}$$

$$\therefore \text{ as } n_A \to \infty, \quad \text{var}\left(\bar{Y}_A\right) \to 0$$

and together with unbiasedness, $\bar{Y}_A \to \mu_A$ (consistent estimator)

Can also show

$$S_A^2 \to \sigma_A^2$$

$$\frac{\#\left\{Y_{iA} \leq x\right\}}{n_A} = \hat{F}_A(x) \to F_A(x) = \int_{-\infty}^{x} p_A(y)dy$$

Connection to Hypothesis Testing

We can then formulate H_0 and H_1 in terms of population quantities:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

Assume:

$$Y_{1A}, \ldots, Y_{n_A A} \sim P_A$$

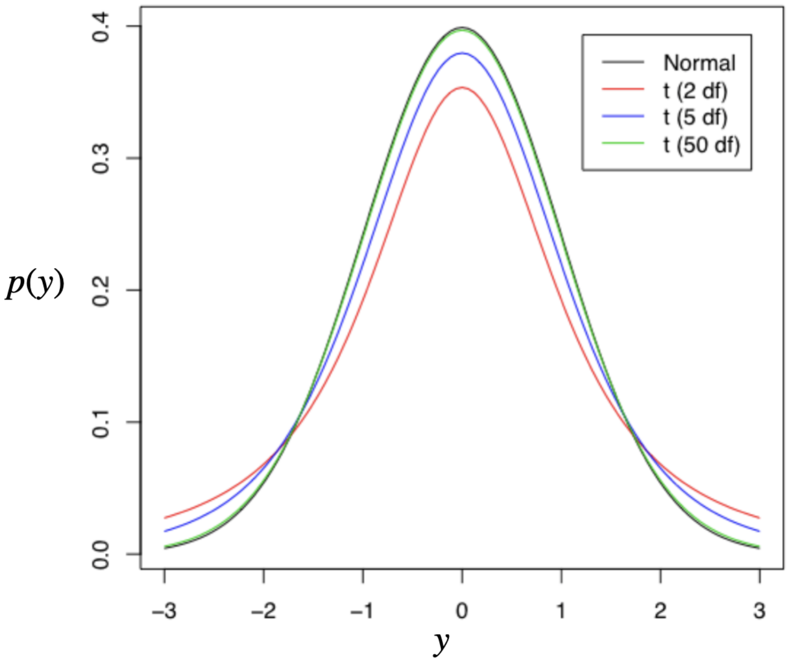
$$Y_{1B}, \ldots, Y_{n_B B} \sim P_B$$

The null hypothesis H_0 can be re-expressed as:

$$\int y p_A(y)\, dy = \int y p_B(y)\, dy$$

To evaluate and get a p-value, we need the distribution of $g(Y_A, Y_B)$ under $\mu_A = \mu_B$. This involves assumptions about P_A and P_B.

The Normal Distribution



Why is this a useful assumption to make?

- Data can be (approximately) normally distributed.
- Sample means are (approximately) normally distributed.

Both due to Central Limit Theorem.
Let $P(\mu, \sigma^2)$ denote a population with mean μ and variance σ^2. Then,

$$\left.\begin{array}{l} Y_1 \sim P_1(\mu_1, \sigma_1^2) \\ Y_2 \sim P_2(\mu_2, \sigma_2^2) \\ \vdots \\ Y_m \sim P_m(\mu_m, \sigma_m^2) \end{array}\right\} \Rightarrow \sum_{j=1}^{m} Y_j \mathrel{\dot\sim} \mathcal{N}\left(\sum_{j=1}^{m} \mu_j, \sum_{j=1}^{m} \sigma_j^2\right)$$

(if the Y_j's are independent)

Sums of varying quantities are approximately normally distributed.

Normally Distributed Data

Consider:

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$$

Empirical distribution of Y_i's will be approximately $N(\mu, \sigma^2)$ where μ and σ^2 depend on $\beta_1, \beta_2, \ldots$, and the means and variances of $X_1, X_2, X_3, \ldots$.

Additive effects $\Rightarrow$ approx normally distributed data.

Normally Distributed Means

For experiment 1:

$$\text{sample: } y_1^{(1)}, \ldots, y_n^{(1)} \text{ i.i.d. } P \Rightarrow \bar{y}^{(1)}$$

$$\vdots$$

For experiment m:

$$\text{sample: } y_1^{(m)}, \ldots, y_n^{(m)} \text{ i.i.d. } P \Rightarrow \bar{y}^{(m)}$$

A histogram of $\{\bar{y}^{(1)}, \ldots, \bar{y}^{(m)}\}$ will be approx $N(\mu, \sigma^2/n)$.

This is the sampling distribution of the sample mean even if the original data are not normal.

Basic Properties of the Normal Distribution

- If $Y \sim N(\mu, \sigma^2)$, then $aY + b \sim N(a\mu + b, a^2\sigma^2)$ for some constants a, b.
- If $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$, and Y_1, Y_2 are independent:

$$\Rightarrow Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- If $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, then $\bar{Y}$ is independent of S^2 (sample variance).

How does this help with hypothesis testing?

Consider:

$$H_0 : \mu_A = \mu_B$$

Under H_0:

$$\bar{Y}_A \dot{\sim} N(\mu, \sigma_A^2/n_A)$$

$$\bar{Y}_B \dot{\sim} N(\mu, \sigma_B^2/n_B)$$

$$\bar{Y}_B - \bar{Y}_A \dot{\sim} N(0, \sigma_{AB}^2)$$

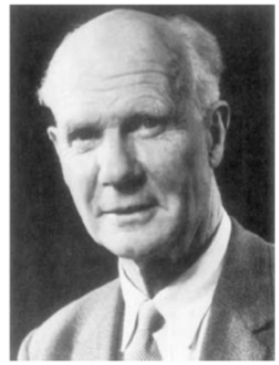
Where:

$$\sigma_{AB}^2 = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$$

$\Rightarrow \therefore$ if know variances, would have a null sampling distribution.

History of the Neyman-Pearson Approach and the p-Value Approach

Jerzy Neyman (1894-1981), Berkeley



Egon Pearson (1895-1980), UCL

Ronald A. Fisher (1890-1962), UCL

The rejection/do not reject dichotomy is associated with the Neyman-Pearson approach to hypothesis testing; p-value is associated with R.A. Fisher.

The t-test (One-sample)

Let's consider first a simple one sample hypothesis test:

$$Y_1, \ldots, Y_n \sim \text{ i.i.d. with mean } \mu \text{ and variance } \sigma^2$$

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Example:

$$Y_i = \text{muscle strength after treatment} - \text{muscle strength before treatment}$$

$$H_0 : \mathbb{E}(Y_i) = \mu = 0$$

Simple and Composite Hypotheses

- The specification of null and alternative hypotheses depends on the problem.

Example

To test whether the mean package weight of a ready-to-eat cereal is 16 ounces, we can set our null hypothesis as

$$H_0 : \mu = 16,$$

and the alternative hypothesis can be

$$H_1 : \mu > 16 \text{ or } H_1 : \mu \neq 16.$$

If the company wants to avoid legal action and/or customer dissatisfaction, then it can set $H_0 : \mu \leq 16$ vs. $H_1 : \mu > 16$.

- The hypothesis like $\mu = 16$, which specifies a single value of μ, is called the simple hypothesis.
- Either of the two alternative hypotheses in the above example includes more than one values of μ, so is called the composite hypothesis.
- Among which, $\mu > 16$ is a one-sided (composite) hypothesis, and $\mu \neq 16$ is a two-sided (composite) hypothesis.

Consider:

$\left|\bar{Y} - \mu_0\right|$ as a test statistic

- sensitive to deviations from H_0
- sampling distribution approximately known

$$\mathbb{E}(\bar{Y}) = \mu$$

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{n}$$

$$\bar{Y} \approx \text{normal}$$

$\therefore$ Under H_0,

$$\left(\bar{Y} - \mu_0\right) \dot{\sim} N\left(0, \frac{\sigma^2}{n}\right)$$

$$\text{but} \quad \sigma^2 \text{ is unknown}$$

What if we scale $\left(\overline{Y} - \mu_0\right)$?

$$\Rightarrow \frac{\overline{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \text{ under } H_0$$

$$\text{where} \quad SE = \sqrt{\text{var}}$$

Having observed data,

$\overline{Y}$ is computable,

n is known,

μ_0 is hypothesized (known),

σ is unknown

Solution: plug in s^2 for σ^2

One Sample t-statistic

For random variable Y, the test statistic is:

$$t(Y) = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

What is the null distribution for $t(\mathbf{y})$?

If approximation $s^2 \approx \sigma^2$ is poor, then we need to take into account uncertainty in our estimate of σ^2.

This can happen for instance with small n.

Some Facts

- If $Z_1, \ldots, Z_n \sim i.i.d \quad N(0,1)$,

$$\Rightarrow \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$$

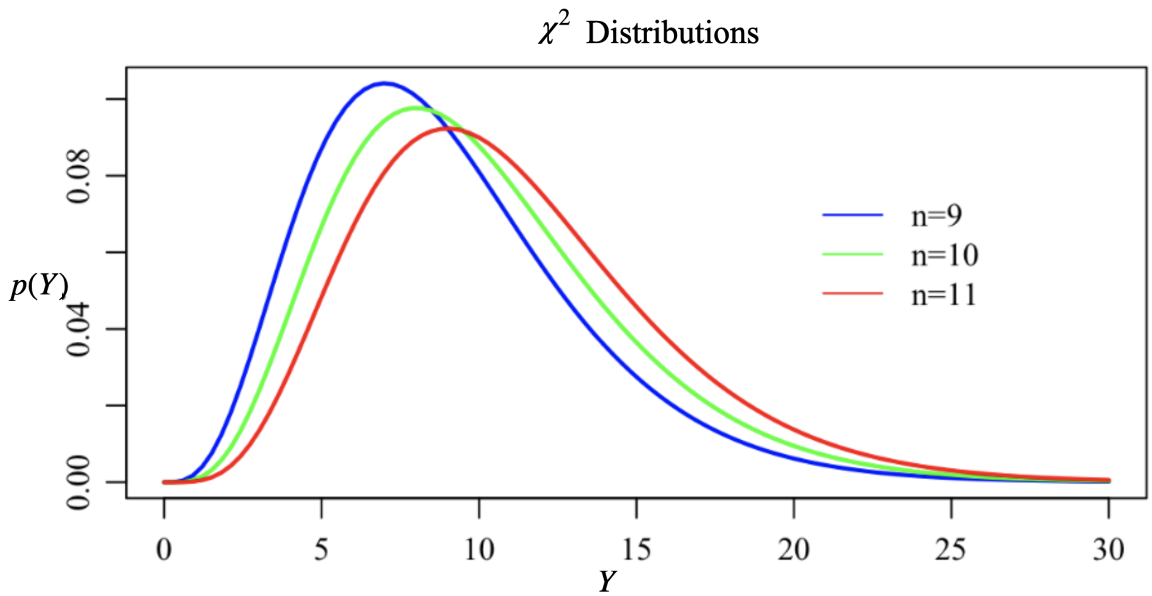
$$also \quad \Rightarrow \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \sim \chi_{n-1}^2$$

- If $Y_1, \ldots, Y_n \quad \text{i.i.d.} \sim \quad \mathcal{N}(\mu, \sigma^2)$

$$\Rightarrow \frac{Y_1 - \mu}{\sigma}, \ldots, \frac{Y_n - \mu}{\sigma} \quad \text{i.i.d.} \sim \quad \mathcal{N}(0,1)$$

$$\Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \sim \chi_n^2$$

$$\text{Also} \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sim \chi_{n-1}^2$$



So all of this allows us to do:

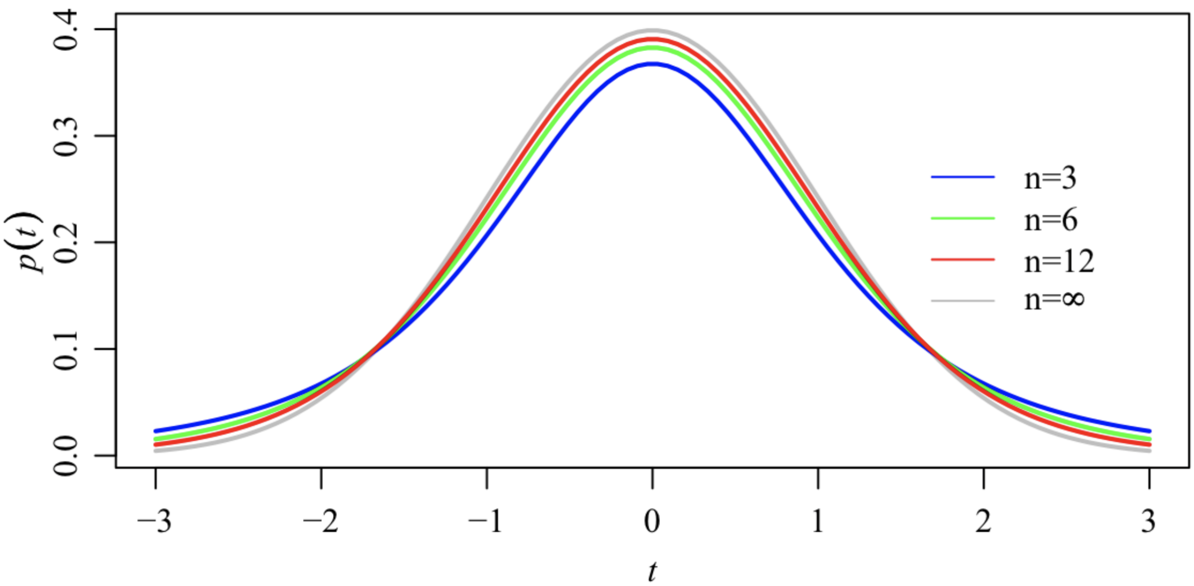
$$\frac{n-1}{\sigma^2}s^2 = \left(\frac{n-1}{\sigma^2}\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right) \sim \chi^2_{n-1}$$

The t-distribution

If:

$$Z \sim N(0,1), \quad X \sim \chi^2_m, \quad Z \text{ and } X \text{ are independent},$$
$$\Rightarrow \frac{Z}{\sqrt{X/m}} \sim t_m$$

See more examples below.



How does this help?

$$\frac{\sqrt{n}\left(\bar{Y}-\mu\right)}{\sigma} \sim \mathcal{N}(0,1)$$

$$\frac{n-1}{\sigma^2}s^2 \sim \chi^2_{n-1}$$

$$\bar{Y} \text{ and } s^2 \text{ are independent}$$

Let

$$Z = \frac{\sqrt{n}\left(\overline{Y}-\mu\right)}{\sigma}$$

$$X = \frac{n-1}{\sigma^2}s^2$$

Then,

$$\frac{Z}{\sqrt{\frac{X}{n-1}}} = \frac{\frac{\sqrt{n}\left(\overline{Y}-\mu\right)}{\sigma}}{\sqrt{\frac{\frac{n-1}{\sigma^2}s^2}{n-1}}} = \frac{\overline{Y}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Under a particular $H_0 : \mu = \mu_0$, this becomes a statistic with known sampling distribution.

Two-sided H_1 and One-sample t-test

$t(Y)$ is the t-statistic as a random variable, $t(y)$ is a observed value.

$$\text{p-value} = P\left(|t(Y)| \geq |t(y)| \mid H_0\right)$$

$$= P(|T_{n-1}| \geq |t(y)|)$$

$$= 2 \cdot P(T_{n-1} \geq |t(y)|)$$

$$\left.\begin{array}{l} = 2 \cdot (1 - pt(t_{obs}, n-1)) \\ = \texttt{t.test(y, mu = mu0)} \end{array}\right\} \textit{Rcode}$$

Level α decision procedure:

Reject H_0 if:

$$\text{p-value} \leq \alpha$$

or equivalently:

$$|t(y)| \geq t_{(n-1),1-\alpha/2}$$

$t_{(n-1),1-\alpha/2}$ is critical value for this test.

for $\alpha = 0.5, \Rightarrow t \approx 2$

History of the t Test

William S. Gosset (1876-1937)

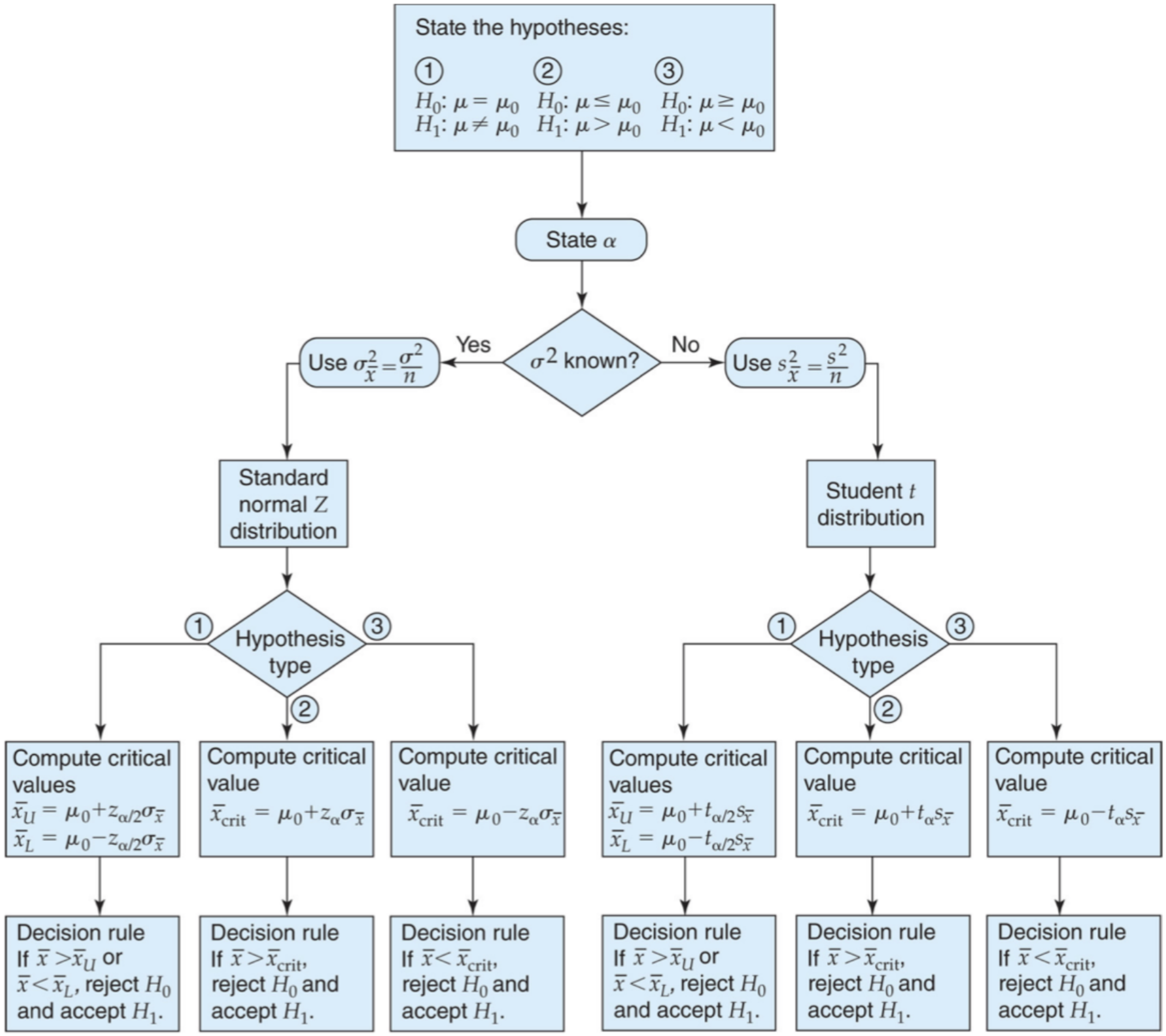
- The t-test is named after Gosset (1908), "The probable error of a mean". At the time, Gosset worked at Guiness Brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym "Student". Consequently, this famous distribution is known as the Student's t rather than Gosset's t! The name "t" was popularized by R.A. Fisher.

Composite Null and Alternative Hypotheses

- $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$
- The data, test statistic, decision rule and p-value are exactly the same as in the previous test.

- $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$
- The data and test statistic are exactly the same as in the previous test.

Schematic Summary



Matched Pair: Two Means

- Matched pair is a kind of dependent samples; apart from the factor under study, the pairs should resemble one another as closely as possible, such as twins.
  - Dependent samples can also be two measurements taken on the same person or object, e.g., a measurement is taken before an event and one after the event (e.g., the treatment on a patient), namely, repeated measurements.
- Data: $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i - y_i \sim N\left(\mu_x - \mu_y, \sigma_d^2\right)$ but x_i and y_i need not be normally distributed, and μ_x, μ_y and σ_d^2 are unknown.
- (i) $H_0 : \mu_x - \mu_y = 0$ or $H_0 : \mu_x - \mu_y \leq 0$ vs. $H_1 : \mu_x - \mu_y > 0$
- (ii) $H_0 : \mu_x - \mu_y = 0$ or $H_0 : \mu_x - \mu_y \geq 0$ vs. $H_1 : \mu_x - \mu_y < 0$
- (ii) $H_0 : \mu_x - \mu_y = 0$ vs. $H_1 : \mu_x - \mu_y \neq 0$
- This is like testing one normal mean with unknown population variance.
  - x_i, μ, μ_0 and σ^2 there are like $x_i - y_i$, $\mu_x - \mu_y$, 0 and σ_d^2 here.

- Test Statistic:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}},$$

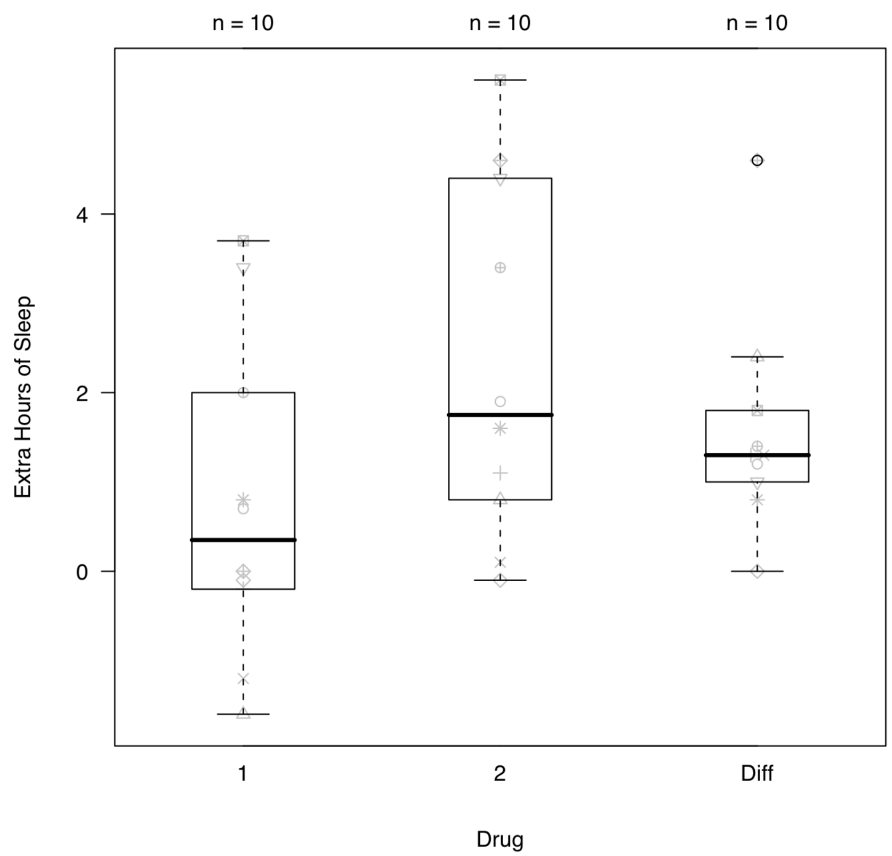
  which follows the t_{n-1} distribution under H_0, where $\bar{d} = \bar{x} - \bar{y}$, and s_d is the sample standard deviation of $\{(x_i - y_i)\}_{i=1}^{n}$.
- Decision Rule: reject H_0 if $t > t_{n-1,\alpha}$ in (i), if $t < -t_{n-1,\alpha}$ in (ii), and $|t| > t_{n-1,\alpha/2}$ in (iii).
- The p-value is $P(T > t)$ in (i), $P(T < t)$ in (ii), and $P(|T| > |t|)$ in (iii), where $T \sim t_{n-1}$.
- (*) Recall that the power of the t-test is inversely affected by σ_d^2, so a smaller σ_d^2 is favorable to the detection of the difference in μ_x and μ_y. Since

$$\sigma_d^2 = Var(x - y) = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy},$$

Example: Paired t-test

- Compare the effects of two soporific drugs.
- Each subject receives placebo, Drug 1, and Drug 2.
- Dependent variable: Number of hours of increased sleep.
- Drug 1 given to n subjects, Drug 2 given to the same n subjects.
- Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
  - $H_0 : \mu_d = 0 \quad$ where $\mu_d = \mu_1 - \mu_2$
  - $H_1 : \mu_d \neq 0$

Subject	**Drug 1**	**Drug 2**	**Diff (2-1)**
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-0.2	1.1	1.3
4	-1.2	0.1	1.3
5	-0.1	-0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4
Mean	0.75	2.33	1.58
SD	1.79	2.0	1.2



- Stat program output (R)

Syntax:

result <- t.test(before, after, paired=T)

```
        Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

- Interpretation
  - A person who takes Drug 2 sleeps on average 1.58 hours longer (95% CI: [0.70, 2.50]) than a person who takes Drug 1

Discuss soon

Two Sample t-test

Sampling model:

$$Y_{1A}, \ldots, Y_{n_A A} \sim \text{i.i.d. } \mathcal{N}(\mu_A, \sigma^2)$$

$$Y_{1B}, \ldots, Y_{n_B B} \sim \text{i.i.d. } \mathcal{N}(\mu_B, \sigma^2)$$

note that:σ^2equal variances assumed

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

Recall:

$$\overline{Y}_B - \overline{Y}_A \sim N(\mu_B - \mu_A, \sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right))$$

If H_0 is true:

$$\overline{Y}_B - \overline{Y}_A \sim N(0, \sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right))$$

How to get plug-in estimates for σ^2?

Consider:

s_p^2: pooled variance

$$s_p^2 = \frac{\sum_{i=1}^{n_A} (Y_i - \bar{Y}_A)^2 + \sum_{i=1}^{n_B} (Y_i - \bar{Y}_B)^2}{(n_A - 1) + (n_B - 1)}$$

$$= \frac{(n_A - 1) s_A^2 + (n_B - 1) s_B^2}{(n_A - 1) + (n_B - 1)}$$

This generates two-sample t-statistic

0 refers to $H_0 : \mu_A = \mu_B$

$$t(Y_A, Y_B) = \frac{(\overline{Y}_B - \overline{Y}_A) - 0}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

$$\overset{H_0}{\sim} t_{n_A+n_B-2}$$

Numerical Example (2 trt data again)

Decision procedure:

- Level α test of $\mathrm{H}_0 : \mu_A = \mu_B$, with $\alpha = 0.05$
- Reject H_0 if p-value < 0.05
- Reject H_0 if $|t(\boldsymbol{y}_A, \boldsymbol{y}_B)| > t_{10,.975} = 2.23$

Data:

- $\bar{y}_A = 18.36,\ s_A^2 = 17.93,\ n_A = 6$
- $\bar{y}_B = 24.30,\ s_B^2 = 26.54,\ n_B = 6$

t-statistic:

- $s_p^2 = 22.24,\ s_p = 4.72$
- $t(y_A, y_B) = 5.93/(4.72\sqrt{1/6 + 1/6}) = 2.18$

Inference:

t.r.v. with 10 df under H_0

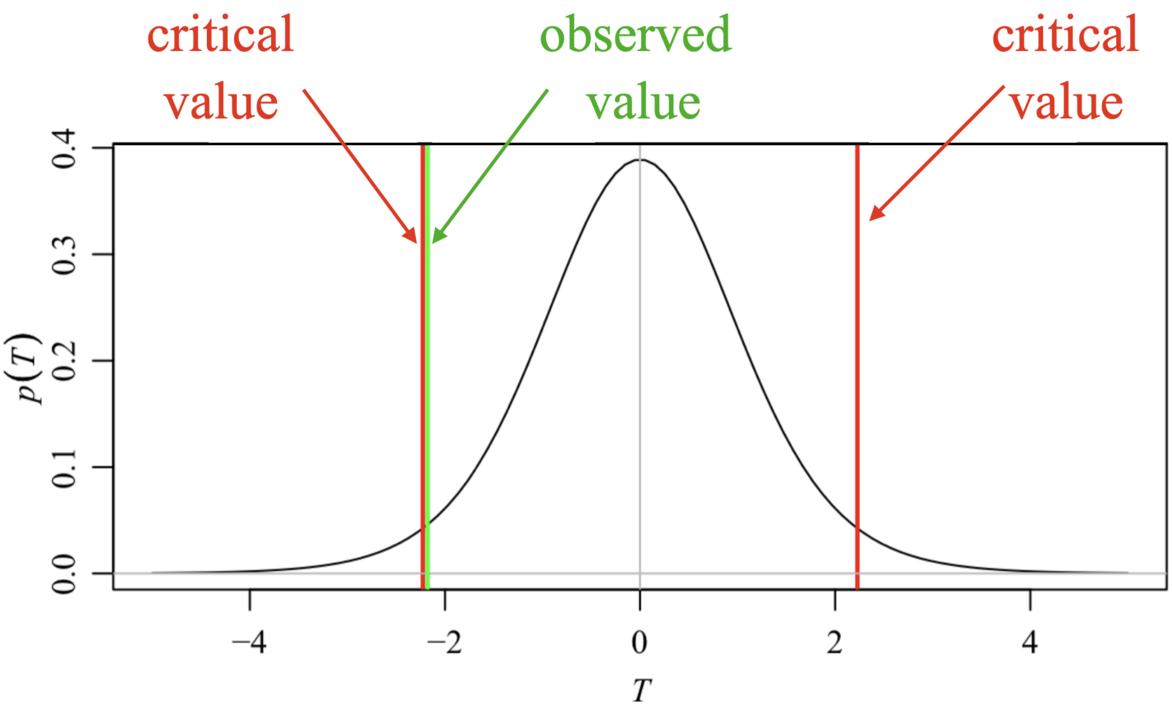
- Hence the p-value$= \Pr(|T_{10}| \geq 2.18) = 0.054$
- Hence H_0: $\mu_A = \mu_B$ is **not rejected** at level $\alpha = 0.05$

```
> t.test(y[x=="A"],y[x=="B"],var.equal=TRUE)

        Two Sample t-test

data:  y[x == "A"] and y[x == "B"]
t = -2.1793, df = 10, p-value = 0.05431
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.999621   0.132954
sample estimates:
mean of x mean of y
 18.36667  24.30000
```

$\boxed{\text{Observed value (in this case } \bar{y}_A - \bar{y}_B \text{)}}$



pdf for a T_{10} (under H_0)

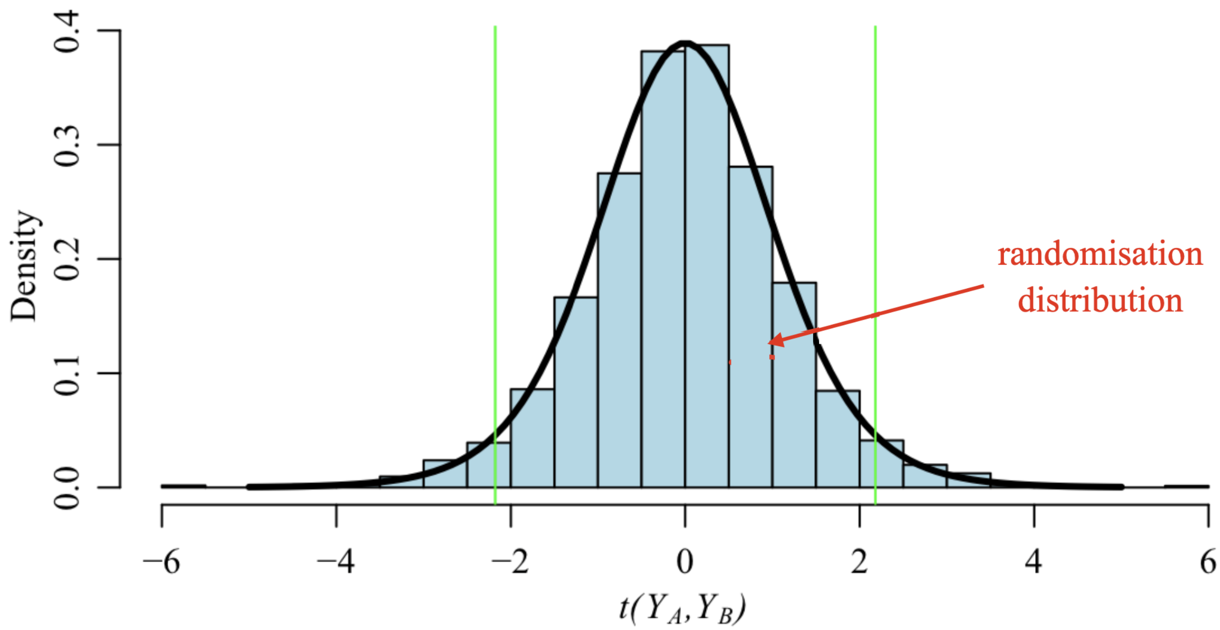
Comparison to Randomization Test

```
t.stat.obs<-t.test( y[x=="A"],y[x=="B"], var.equal=T)$stat
t.stat.sim<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  tmp<-t.test(y[xsim=="B"],y[xsim=="A"],var.equal=T)
  t.stat.sim[s]<-tmp$stat
}

mean( abs(t.stat.sim) >= abs(t.stat.obs) )
```

=0.058

(p-value from 2 sample t-test = 0.054)



Is there concordance between the two approaches surprising?
Assumptions: Under H_0

- **Randomization Test:**
  1. Treatments are randomly assigned
- **t-test:**
  1. Data are independent samples
  2. Each population is normally distributed
  3. The two populations have the same variance

Imagined Universes:

- Randomization Test: **Numerical responses** remain **fixed**, we imagine only alternative treatment assignments.
- t-test: **Treatment assignments** remain **fixed**, we imagine an alternative sample of experimental units and/or conditions, giving **different numerical responses**.

Inferential Context / Type of Generalization

- Randomization Test: Inference is specific to our particular experimental units and conditions.
- t-test: Under our assumptions, inference claims to be **generalizable** to other units / conditions, i.e., to a larger population.

Yet the numerical results are often nearly identical.

Keep the following concepts clear:

- **t-statistic**: a scaled difference in sample means, computed from the data.
- **t-distribution**: the probability distribution of a normal random variable divided by the square-root of a χ^2 random variable.
- **t-test**: a comparison of a t-statistic to a t-distribution.
- **randomization distribution**: the probability distribution of a test statistic under random treatment reassignments and H_0.
- **randomization test**: a comparison of a test statistic to its randomization distribution.
- **randomization test with the t-statistic**: a comparison of the t-statistic to its randomization distribution.

Checking Assumptions

Two sample t-test assumes:

a) $\mu_A = \mu_B$ (to derive sampling distribution under H_0)

b) $\sigma_A^2 = \sigma_B^2$

c) p_A and p_B are normal distributions

Thus rejecting $H_0 \Rightarrow a)$ is not true,
$\therefore$ want to check $b)$ and $c)$.

Checking Normality

- Use normal probability plot.
- Idea: order observed observations within each group.

$$Y_{(1)A}, \cdots, Y_{(n_A)A}; \quad Y_{(1)B}, \cdots, Y_{(n_B)B}$$

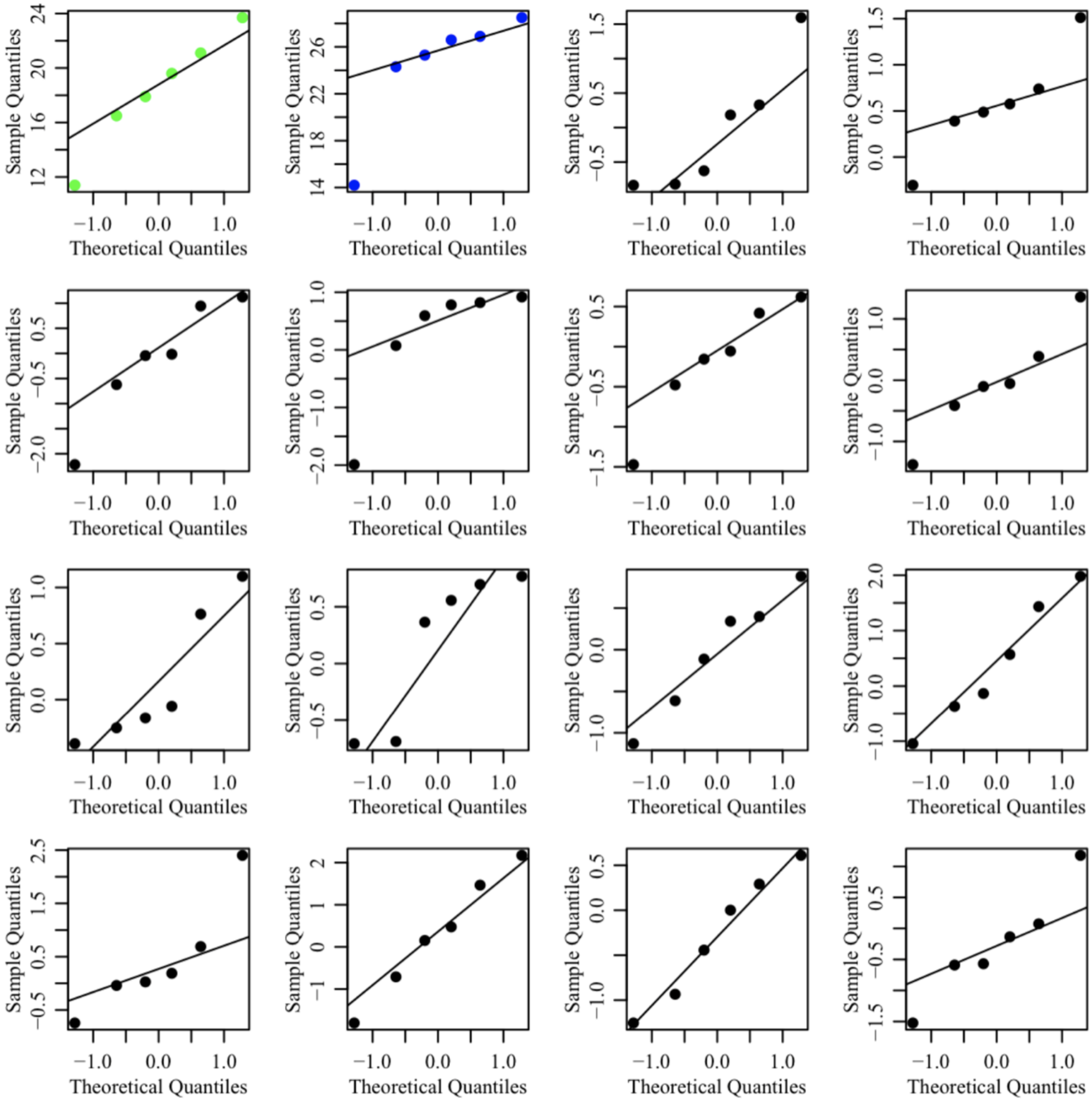
Compare these sample quantiles to theoretical quantiles of a normal distribution ($N(0,1)$ for convenience).

$$Z_{\left(\frac{1}{n_A}-\frac{1}{2}\right)} \leq Z_{\left(\frac{2}{n_A}-\frac{1}{2}\right)} \leq \cdots \leq Z_{\left(\frac{k}{n_A}-\frac{1}{2}\right)}$$

$$\text{where} \quad P\left(Z \leq Z_{\left(\frac{k}{n_A}-\frac{1}{2}\right)}\right) = \frac{k}{n_A} - \frac{1}{2}$$

[and similarly for the Y_{iB} observations]

Then plot one against the other.
If normality holds, the relationship should be approximately linear.



Examples of normal probability plots

Checking Equality of Variances

There is a formal procedure (e.g., Levene's test). We will see this later.
For now, we can use the following rule of thumb:

$$\frac{1}{4} < \frac{S_A^2}{S_B^2} < 4 \quad \text{(A, B can reverse)}$$

What if variances do not appear to be equal?

Options:

1. Use randomization H_0 distribution.

2. Transform data to stabilize variances.

3. Use modified t-test that allows unequal variances:

$$t_W(Y_A, Y_B) = \frac{\overline{Y}_B - \overline{Y}_A}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

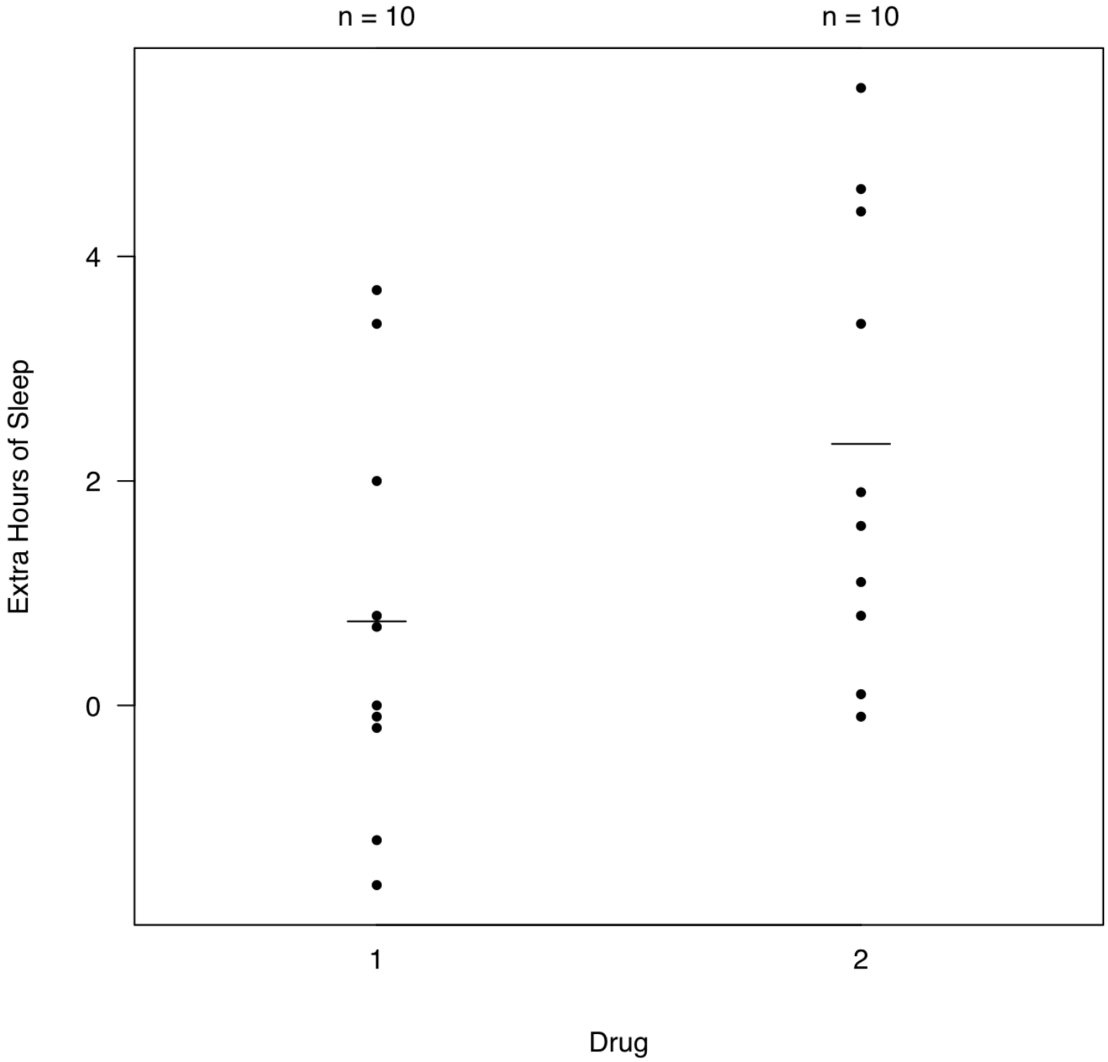
W: df based on Welch's approximation

Example: Two-Sample t-test

- Compare the effects of two soporific drugs
  - Optical isomers of hyoscyamine hydrobromide
- Each subject receives a placebo and then is randomly assigned to receive Drug 1 or Drug 2
- Dependent variable: Number of hours of increased sleep over control
- Drug 1 given to n_1 subjects, Drug 2 given to n_2 different subjects
- Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

Obs.	**Drug 1**	**Drug 2**
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4
Mean	0.75	2.33
SD	1.79	2.0


n = 10
n = 10
4
2
0
Extra Hours of Sleep
1
2
Drug

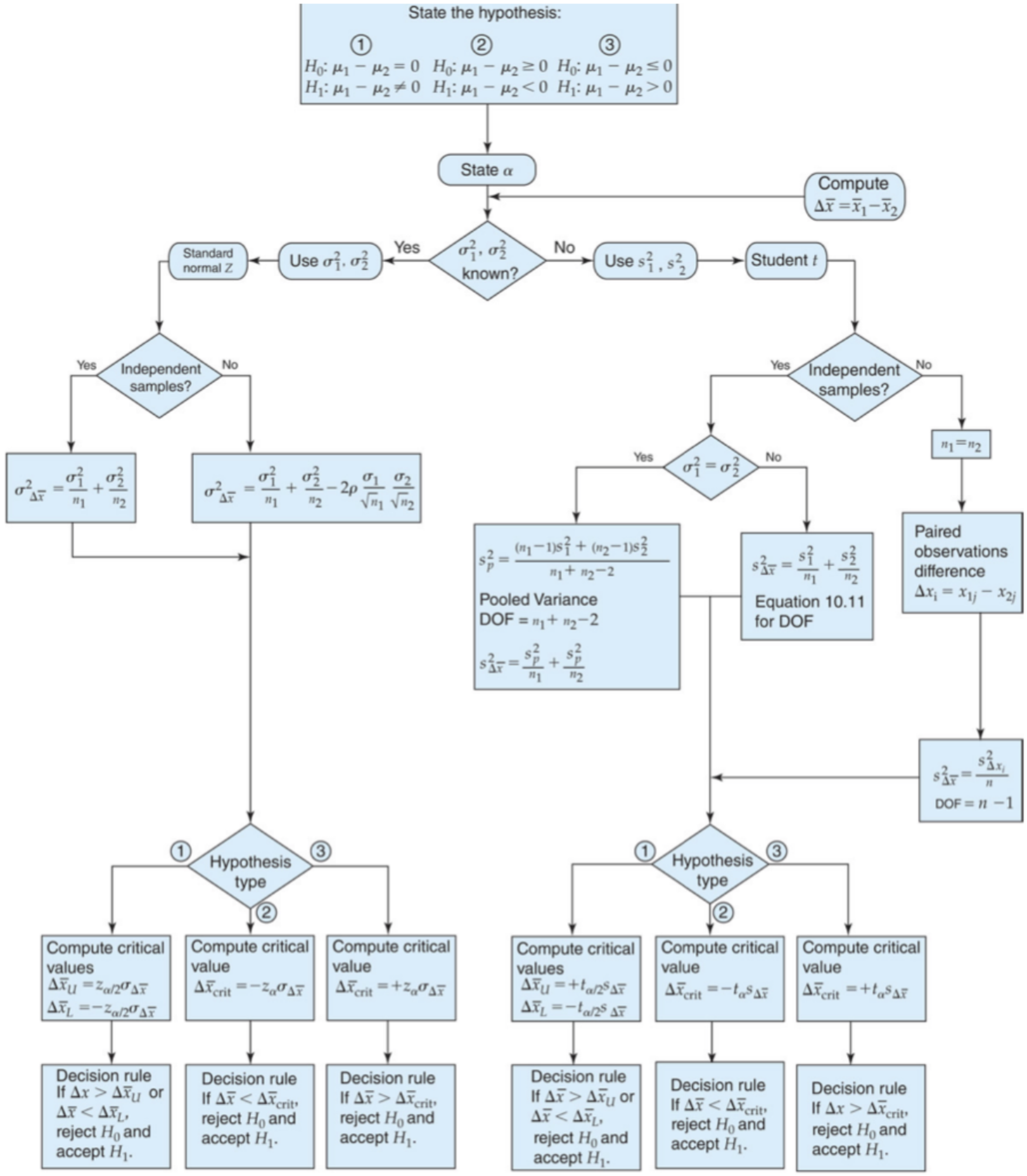
- Stat program output $\boxed{\text{(R)}}$

```
        Two Sample t-test
```

$\boxed{\text{Syntax: See earlier notes}}$

```
data:  extra by group
t = -1.8608, df = 18, p-value = 0.07919
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3638740  0.2038740
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

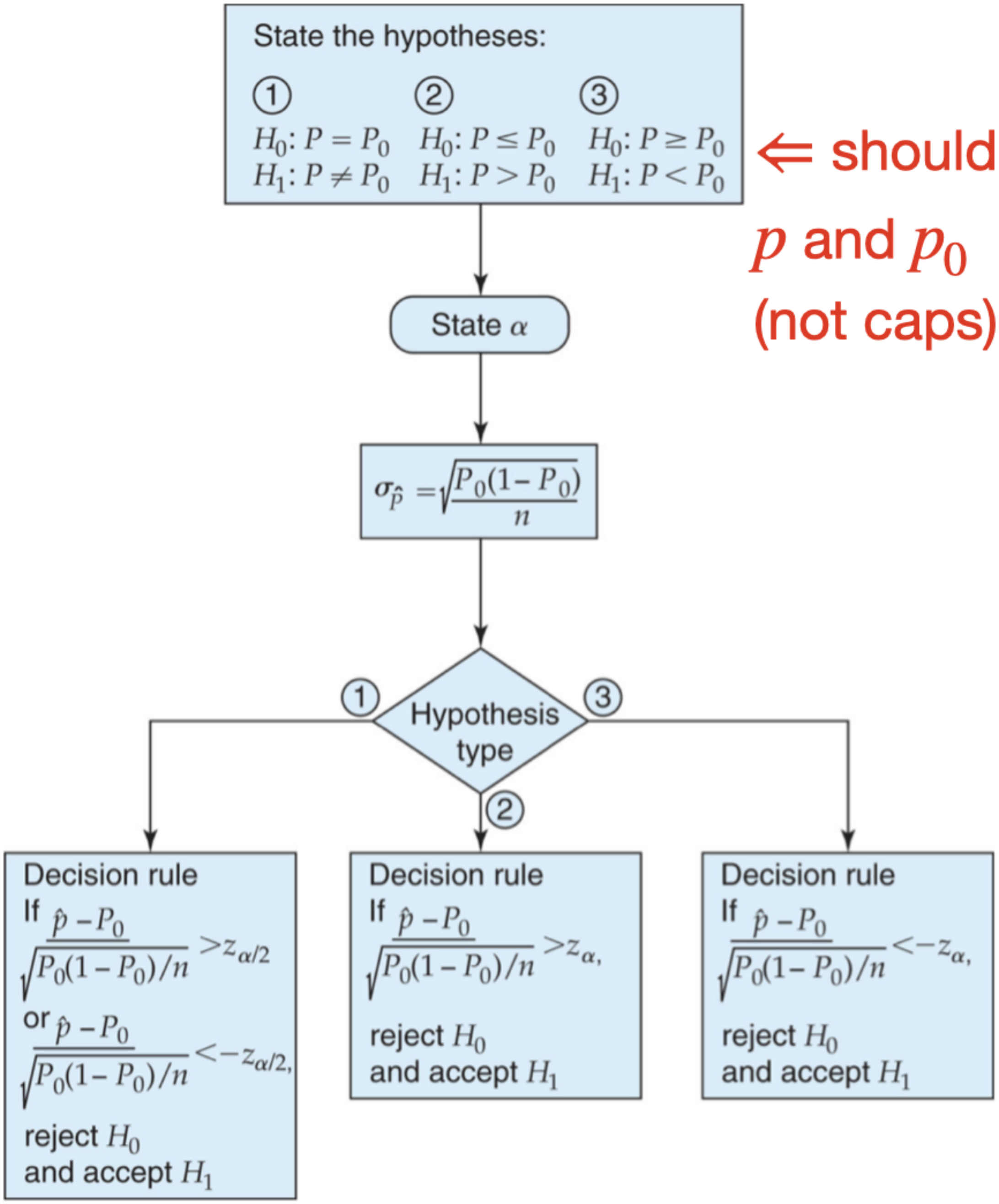
- Interpretation
  - Compare Drug 2 to Drug 1. The output compares 1 to 2
  - Individuals who take Drug 2 sleep on average 1.58 hours longer (95% CI: [-0.20, 3.36]) than individuals who take Drug 1



State the hypothesis:
① $H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$
② $H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$
③ $H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$
State α
Compute $\Delta\bar{x} = \bar{x}_1 - \bar{x}_2$
σ_1^2, σ_2^2 known?
Yes
No
Use σ_1^2, σ_2^2
Standard normal Z
Use s_1^2, s_2^2
Student t
Independent samples?
Yes
No
$\sigma^2_{\Delta\bar{x}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
$\sigma^2_{\Delta\bar{x}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2\rho \frac{\sigma_1}{\sqrt{n_1}} \frac{\sigma_2}{\sqrt{n_2}}$
$\sigma_1^2 = \sigma_2^2$
$n_1 = n_2$
$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
Pooled Variance
DOF = $n_1 + n_2 - 2$
$s^2_{\Delta\bar{x}} = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}$
$s^2_{\Delta\bar{x}} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$
Equation 10.11 for DOF
Paired observations difference
$\Delta x_i = x_{1j} - x_{2j}$
$s^2_{\Delta\bar{x}} = \frac{s^2_{\Delta x_i}}{n}$
DOF = $n - 1$
Hypothesis type
① ② ③
Compute critical values $\Delta\bar{x}_U = z_{\alpha/2}\sigma_{\Delta\bar{x}}$ $\Delta\bar{x}_L = -z_{\alpha/2}\sigma_{\Delta\bar{x}}$
Compute critical value $\Delta\bar{x}_{crit} = -z_\alpha\sigma_{\Delta\bar{x}}$
Compute critical value $\Delta\bar{x}_{crit} = +z_\alpha\sigma_{\Delta\bar{x}}$
Compute critical values $\Delta\bar{x}_U = +t_{\alpha/2}s_{\Delta\bar{x}}$ $\Delta\bar{x}_L = -t_{\alpha/2}s_{\Delta\bar{x}}$
Compute critical value $\Delta\bar{x}_{crit} = -t_\alpha s_{\Delta\bar{x}}$
Compute critical value $\Delta\bar{x}_{crit} = +t_\alpha s_{\Delta\bar{x}}$
Decision rule If $\Delta x > \Delta\bar{x}_U$ or $\Delta\bar{x} < \Delta\bar{x}_L$, reject H_0 and accept H_1.
Decision rule If $\Delta\bar{x} < \Delta\bar{x}_{crit}$, reject H_0 and accept H_1.
Decision rule If $\Delta\bar{x} > \Delta\bar{x}_{crit}$, reject H_0 and accept H_1.
Decision rule If $\Delta\bar{x} > \Delta\bar{x}_U$ or $\Delta\bar{x} < \Delta\bar{x}_L$, reject H_0 and accept H_1.
Decision rule If $\Delta\bar{x} < \Delta\bar{x}_{crit}$, reject H_0 and accept H_1.
Decision rule If $\Delta x > \Delta\bar{x}_{crit}$, reject H_0 and accept H_1.

One Proportion, Large Samples

- Data: same as in the previous test, but X_i can only take 0 or 1 and follows the Bernoulli(p) distribution.
- The three pairs of hypotheses are the same as in the previous test, but here the population means are denoted as p.
- Test Statistic:
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 (1 - p_0) / n}},$$
which follows the $N(0, 1)$ distribution under H_0 in large sample [$np_0 (1 - p_0) > 5$ with p_0 being the proportion under H_0], where $\hat{p} = \bar{x}$ is the sample proportion.
- Recall that the variance of the Bernoulli(p) distribution is $p(1 - p)$, so under H_0, the variance of x_i is known. This is like testing one normal mean with known population variance.
In schematic on next slide
- Decision Rule: reject H_0 if $z > z_\alpha$ in(ii), if $z < -z_\alpha$ in(iii), and $|z| > z_{\alpha/2}$ in (i) .
- The p-value formulae are the same as in testing one normal mean with known population variance.



$\Leftarrow$ should be p and p_0 (not caps)

Independent Samples: Two Proportions, Large Samples

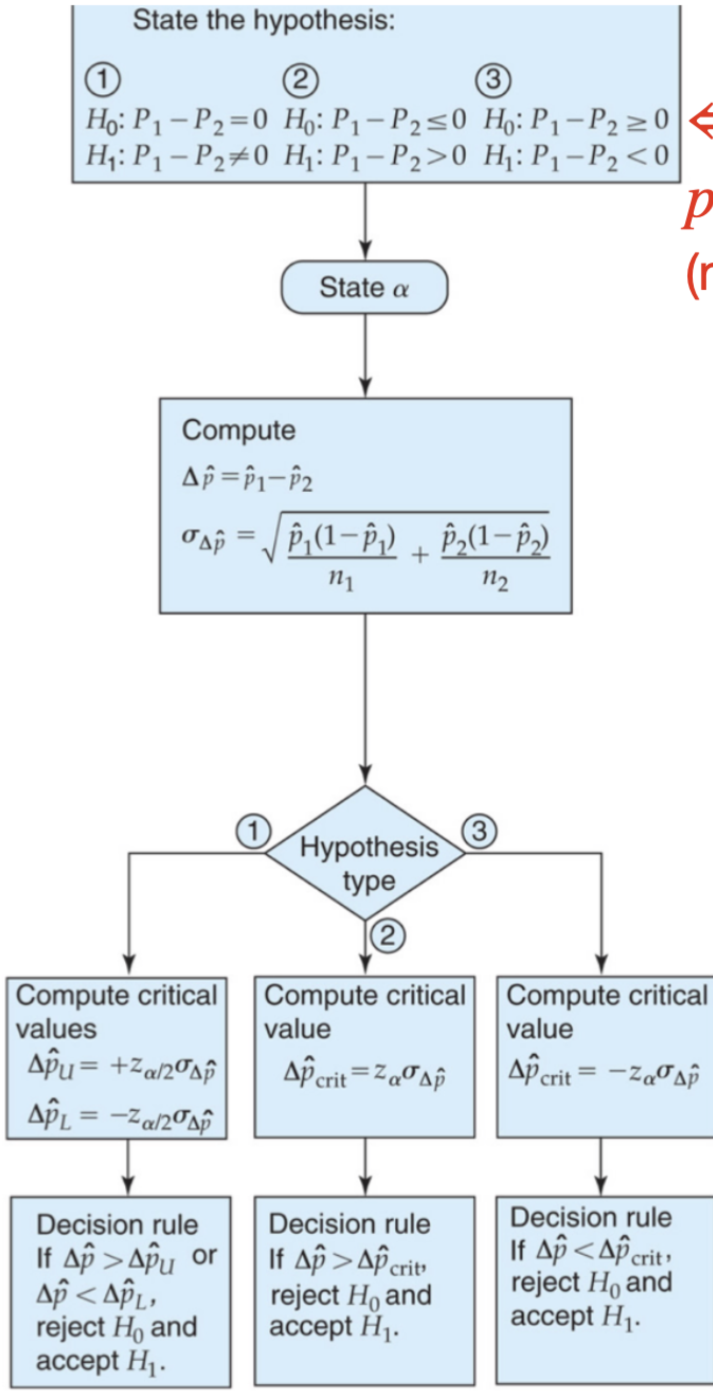
- Data: same as in the previous test, but X_i and Y_i can only take 0 or 1, so follows the Bernoulli(p_x) and Bernoulli(p_y) distributions.
- The three pairs of hypotheses are the same as in the previous test, but here the population means are denoted as p_x and p_y.
- Test Statistic:

$$\text{some use } Z \rightarrow t = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}},$$

  which follows the $N(0,1)$ distribution under H_0 in large sample [$np_0(1-p_0) > 5$ with p_0 being the common proportion under H_0], where

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y} = \frac{\text{total number of successes in } \{x_i\}_{i=1}^{n_x} \cup \{y_j\}_{j=1}^{n_y}}{\text{total sample size of } \{x_i\}_{i=1}^{n_x} \cup \{y_j\}_{j=1}^{n_y}}.$$

  - Recall that the variance of the Bernoulli(p) distribution is $p(1-p)$, so under H_0, the variances of x_i and y_i are also equal. This is like testing two normal means with unknown equal population variances. see schematic before
- Decision Rule: reject H_0 if $t > z_\alpha$ in (ii), if $t < -z_\alpha$ in (iii), and $|t| > z_{\alpha/2}$ in (i) .
- The p-value is $P(Z > t)$ in (i), $P(Z < t)$ in (ii), and $P(|Z| > |t|)$ in (iii), where $Z \sim N(0,1)$.



$\Leftarrow$ should be p_1 and p_2 (not caps)

Example:
Test whether the population of women whose age at first birth ≤ 29 has the same probability of breast cancer as women whose age at first birth was ≥ 30. This dichotomization is highly arbitrary and we should really be testing for an association between age and cancer incidence, treating age as a continuous variable.

- **Case-control study** (independent and dependent variables interchanged); p_1 = probability of age at first birth ≥ 30, etc.

	with Cancer	**without Cancer**
Total # of subjects	3220 (n_1)	10245 (n_2)
# age $\geq$ 30	683	1498
Sample probabilities	0.212 ($\hat{p}_1$)	0.146 ($\hat{p}_2$)

  – Pooled probability:

$$\frac{683 + 1498}{3220 + 10245} = 0.162$$

- **Estimate the variance**

$$\text{variance}(\hat{p}_1 - \hat{p}_2) = \hat{p}_0(1 - \hat{p}_0) \times \left[\frac{1}{n_1} + \frac{1}{n_2}\right] = 5.54 \times 10^{-5}$$

$$SE = \sqrt{\text{variance}} = 0.00744$$

- **Test statistic**

$$z = \frac{0.212 - 0.146}{0.00744} = 8.85$$

- 2-tailed P-value is 0.0 using `survstat`; we report $P < 0.0001$

- We do not use a t-distribution because there is no σ to estimate (and hence no "denominator d.f." to subtract).

Other testing approaches include:

- χ^2 test
- Fisher's exact test