Hypothesis Tests II

Sumanta Basu

Hypothesis testing

Hypothesis testing

Goal: make decisions about a population parameter based on a sample of data.

Statistical hypothesis - a statement made about the value of a population parameter (e.g. $\mu > 80$)

Hypothesis test - statistical method for evaluating the degree to which data favors (or does not favor) the “alternative” hypothesis over the null hypothesis.

Example

Research question: Can I read your minds?

The data

n = Number of people in room

x = Number that I got correct

Is x large enough for us to believe that I'm psychic?

Guiding mindset

> We want to find a simplest explanation of the observed phenomenon

- Unless there is strong enough evidence to the contrary, we should assume that I am random guessing.
- **Thinking like a skeptic:** If I were random guessing, would getting x correct out of n be surprising?

Statistics to the rescue

- x is a realization of what sort of random variable?

$$X \sim \text{Binomial}(n, p)$$

Null hypothesis - expresses skeptical perspective, i.e., "nothing interesting here" (status quo) - in this example - "He's random guessing." - $H_0 : p = 1/4$

Alternative hypothesis - something new, not previously accepted - He's psychic! $H_A : p > 1/4$.

Is this surprising "under the null?"

$$H_0 : p = 1/4$$

Under null, we think x is a realization of random variable

$$X \sim \text{Binomial}(n, 1/4).$$

x is higher than $n/4$. But is it unlikely under random guessing?

If $P(X \geq x)$ is very small under null hypothesis, perhaps we should favor alternative that $p > 1/4$.

In R

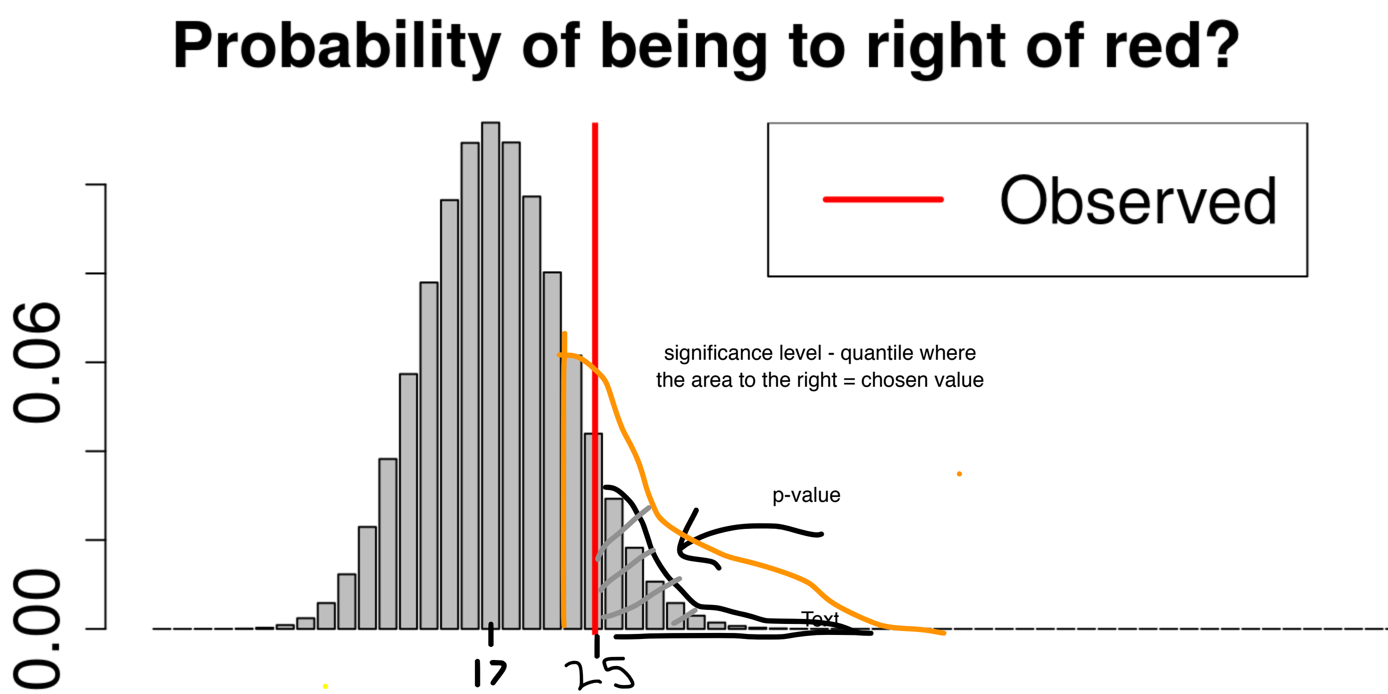
```r
n=65 # number of trials
x=25 # number of successes
p = 1/4 # calculate under null hypothesis
1 - pbinom(x-1, size = n, prob = p)
```

```
## [1] 0.011344
```

If Null were true, what are the odds that I'd see something as or more extreme than what was observed

- This is called a **p-value** - *the probability, calculated under the null, of seeing something as extreme or more extreme than what was observed.*
- Note: direction of "extreme" is defined by alternative hypothesis

In a picture


Probability of being to right of red?
Observed
0.06
0.00
significance level - quantile where
the area to the right = chosen value
p-value
Text
17
25

How small is small enough?

Upper Bound on Making Type I error

- the **significance level** α of a hypothesis test is our chosen threshold for the p-value, below which we reject the null.
- most common choice: 0.05… i.e., 1/20.
- Are you surprised if something happens to you that should only happen 1 out of every 20 times?

> *"A claimed result that overturns all ideas of causality might well require something stricter than .05."* - Brad Efron (NY Times 2011)

Connecting back to science

Your goal: Convince skeptical reader of your “research finding” - Is observed data necessarily inconsistent with a more simple explanation? - **Structure of argument:** - There are two possibilities:

```
1) simple ("null") explanation

2) new ("alternative") science here
```

- Our data would be very unusual if (1) were true.
- We and reader are forced to reject (1) in favor of (2)

Two kinds of errors

	H_0 true	H_A true
Reject H_0	Type I error	Good Sensitivity
Fail to reject H_0	Good Specificity	Type II error

Type I error - false positive ("gullible")

Type II error - false negative ("missed out on an opportunity")

Goal: design a procedure that can ensure that

$$P(\text{Type I error}) \leq \alpha$$

Chance of getting it wrong. e.g. if \alpha = 0.05 we have a 5% chance of being wrong. Upper bound on making a type I error

yet still has small $P(\text{Type II error})$.

Significance level

By $P(\text{Type I error})$ we mean

$$P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

Think of **significance level** as our *level of gullibility*

- thinking I'm psychic when I'm random guessing
- declaring drug works when it actually makes no difference
- jury deciding "guilty" when person is innocent

Power

•1 - P(Type II Error)
•Enough data to capture the signal
•Related to Sensitivity

The **power** of a test is

$$P(\text{Reject } H_0 \mid H_A \text{ is true})$$

Sensitivity

Power is test's **ability to detect** that alternative applies.

- detecting that drug works when it in fact does
- jury deciding "guilty" when person was guilty of crime

Note:

$$P(\text{Type II error}) = P(\text{Fail to reject } H_0 \mid H_A \text{ is true}) = 1 - \text{Power}$$

Back to example

Test

$$H_0 : p = 1/4 \text{ versus } H_A : p > 1/4$$

Suppose I want a test with significance level $\alpha = 0.05$. How high would observed x need to be for me to reject H_0?

Consider decision rule in which I reject H_0 if observed x is $\geq c$.

Want to find a cutoff c such that

Type I error

$$P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(X \geq c \mid H_0 \text{ true}) = 0.05$$

Finding cutoff

$$
\begin{aligned}
P(X \geq c \mid H_0 \text{ true}) &= P(\text{Binomial}(n, 1/4) \geq c) \\
&= 1 - P(\text{Binomial}(n, 1/4) \leq c - 1)
\end{aligned}
$$

```
alpha = 0.05; p = 1/4 # under null
quantile = qbinom(1-alpha, n, p) # this is c - 1
cutoff = quantile + 1
cutoff
```

```
## [1] 23
```

```
1 - pbinom(cutoff - 1, n, p)
```

```
## [1] 0.04024569
```

Rejection region

Our level $\alpha = 0.05$ test rejects if observed number of successes is greater than or equal to 23.

We have designed the **rejection region** of the test (the set of values for which we will reject H_0) so that

$$P(\text{Reject } H_0 \mid H_0 \text{ is true}) \leq 0.05$$

Making the case

Suppose we found that I got $x = 21$ correct out of $n = 65$ guesses.

We fail to reject H_0 because 21 is less than 23.

Have we shown that I am not psychic?

> "Absence of evidence is not evidence of absence."

- Difference between "not guilty" and "innocent"
- Similarly we say "fail to reject H_0" rather than "accept H_0"
- If we fail to reject null, it could just mean we didn't get enough data ("under-powered study")

Power

Suppose I were *slightly* psychic:

$$p = 1/4 + 0.01 = 0.26$$

What's the probability under this alternative that our test would have rejected the null at the $\alpha = 0.05$ level?

Recall: our test rejects H_0 if we observe ≥ 23 successes out of $n = 65$.

$$P(\text{Reject } H_0 \mid p = 0.26) = P(X \geq 23 \mid p = 0.26)$$

where $X \sim \text{Binomial}(n, p)$.

Significance level calculates type 1 error with no power

Power

$$\text{Power} = P\left(\text{Binomial}(65, 0.26) \geq 23\right) = ?$$

```r
cutoff
```

```
## [1] 23
```

```r
1 - pbinom(cutoff - 1, n, prob = 0.26)
```

```
## [1] 0.05988829
```

This is very low power, meaning that if $p = 0.26$, my experiment had very little shot at establishing this.

Power

Power is a function of sample size

Suppose I am *very* psychic, so that $p = 1/2$.

$$\text{Power} = P\left(\text{Binomial}(65, 0.5) \geq 23\right) = ?$$

```
cutoff
```

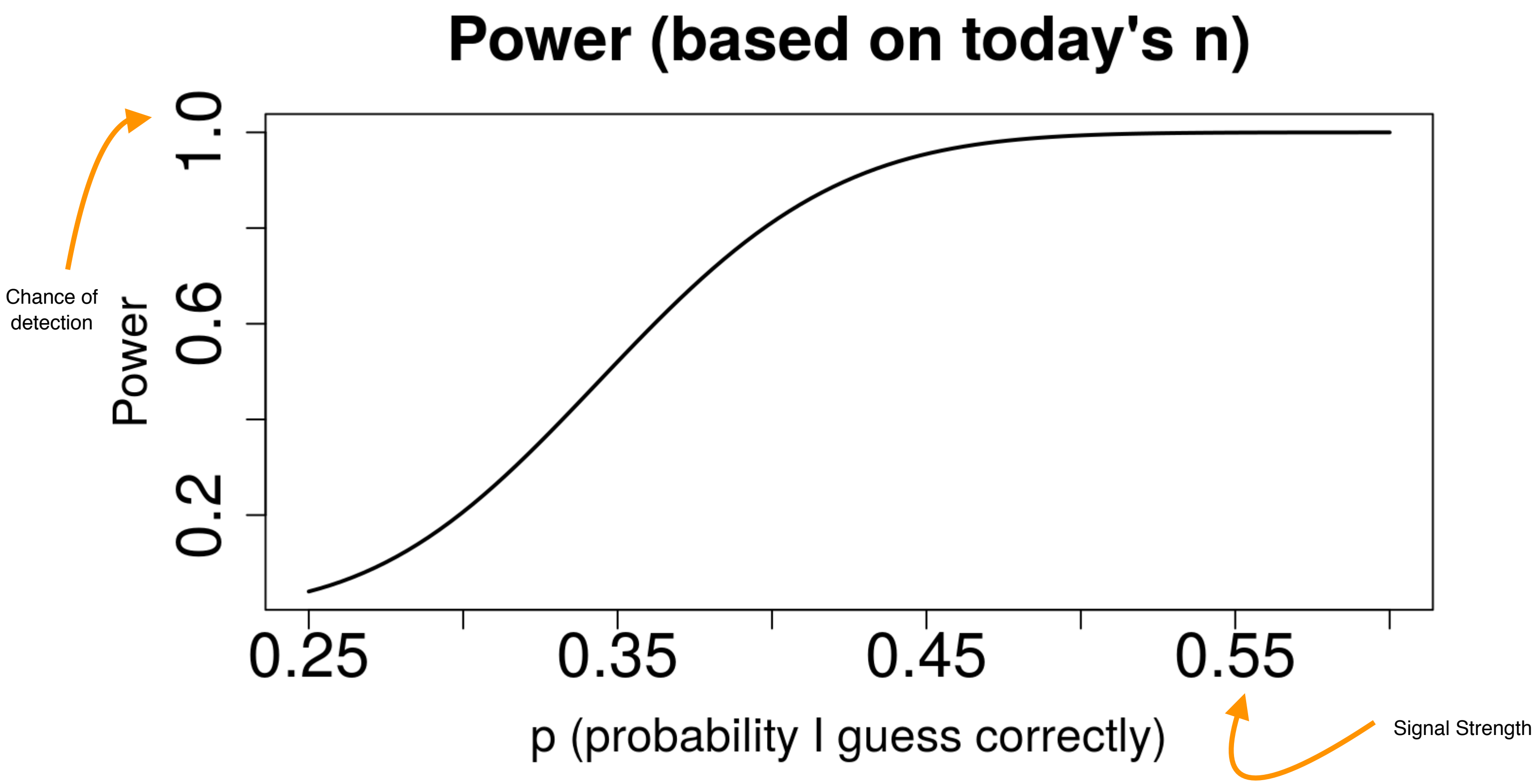
```
## [1] 23
```

```
1 - pbinom(cutoff - 1, n, prob = 0.5)
```

```
## [1] 0.9937487
```

This is very high power, meaning that if $p = 0.5$, my experiment had a very good chance of detecting my abilities.

Power function



A power calculation

Idea: Before doing an experiment, I should figure out what size sample is needed to have a target power.

Requires that I have a guess of the size of p.

A power calculation

Suppose I think I'm slightly psychic: $p = 1/4 + 0.01 = 0.26$. **What n do I need to have 85% power?**

Is $n = 1000$ enough?

```
n = 1000 # initial guess
alpha = 0.05; pnull = 1/4 # under null
quantile = qbinom(1-alpha, n, pnull) # this is c - 1
cutoff = quantile + 1 # this is the cutoff to ensure a level alpha test
palt = 0.26 # suppose I think I'm slightly psychic: 1/4 + 0.01
power = 1 - pbinom(cutoff - 1, n, prob = palt)
power
```

```
## [1] 0.165125
```

A power calculation

Is $n = 2000$ enough?

```
n = 2000 # initial guess
alpha = 0.05; pnull = 1/4 # under null
quantile = qbinom(1-alpha, n, pnull) # this is c - 1
cutoff = quantile + 1 # this is the cutoff to ensure a level alpha test
palt = 0.26 # suppose I think I'm slightly psychic: 1/4 + 0.01
power = 1 - pbinom(cutoff - 1, n, prob = palt)
power
```

```
## [1] 0.2612058
```

A power calculation

Is $n = 10000$ enough?

```
n = 10000 # initial guess
alpha = 0.05; pnull = 1/4 # under null
quantile = qbinom(1-alpha, n, pnull) # this is c - 1
cutoff = quantile + 1 # this is the cutoff to ensure a level alpha test
palt = 0.26 # suppose I think I'm slightly psychic: 1/4 + 0.01
power = 1 - pbinom(cutoff - 1, n, prob = palt)
power
```

```
## [1] 0.7417297
```

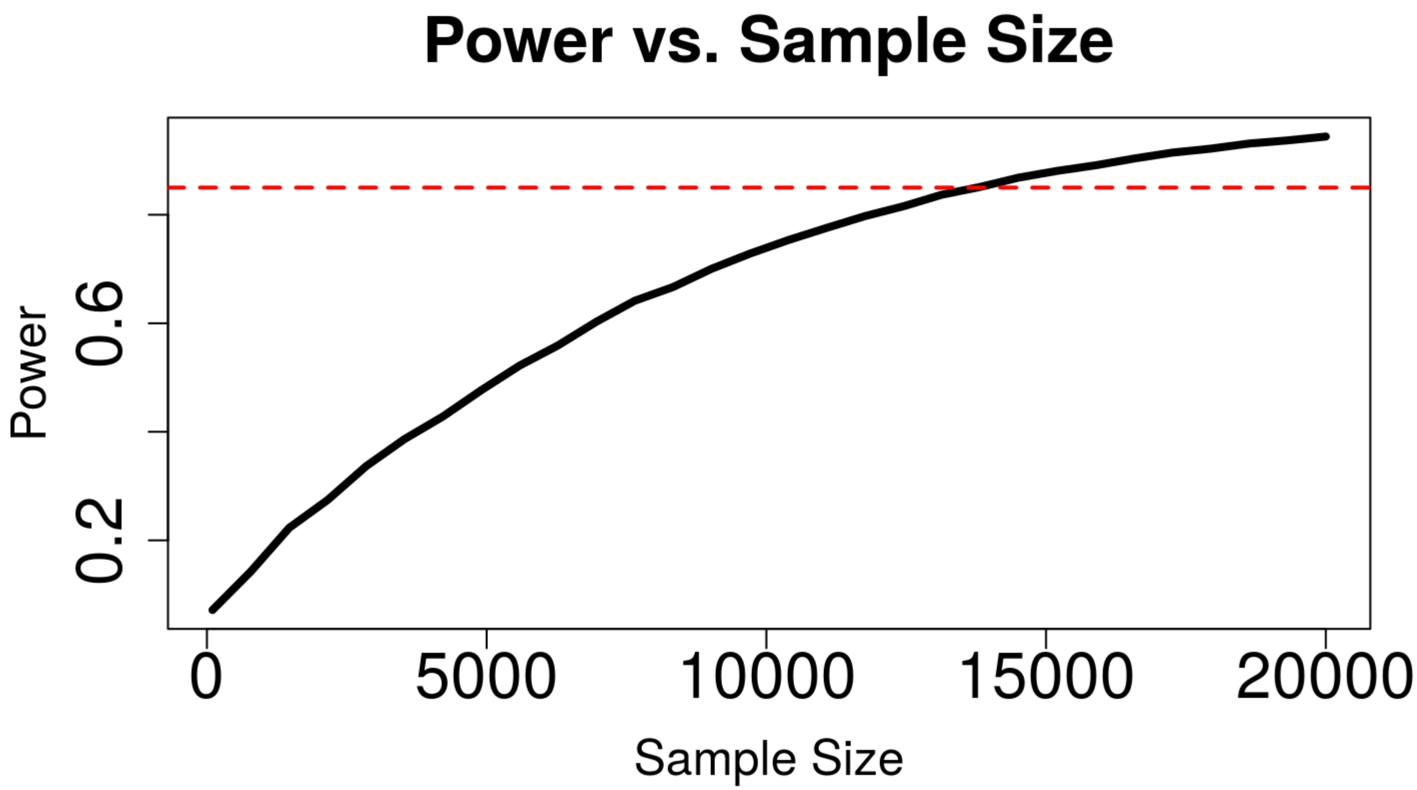
A power calculation

Is $n = 20000$ enough?

```
n = 20000 # initial guess
alpha = 0.05; pnull = 1/4 # under null
quantile = qbinom(1-alpha, n, pnull) # this is c - 1
cutoff = quantile + 1 # this is the cutoff to ensure a level alpha test
palt = 0.26 # suppose I think I'm slightly psychic: 1/4 + 0.01
power = 1 - pbinom(cutoff - 1, n, prob = palt)
power
```

```
## [1] 0.944068
```

Power versus sample size



Power versus sample size

```r
alpha = 0.05; pnull = 1/4 # under null
palt = 0.26 # suppose I think I'm slightly psychic: 1/4 + 0.01
nlist = round(seq(100, 20000, length=50))
power = rep(NA, length(nlist))
for (i in 1:length(nlist)) {
  quantile = qbinom(1-alpha, nlist[i], pnull) # this is c - 1
  cutoff = quantile + 1 # this is the cutoff to ensure a level alpha test
  power[i] = 1 - pbinom(cutoff - 1, nlist[i], prob = palt)
}
```

```r
plot(nlist, power, type="l", xlab="n", ylab="Power", main="Power vs. Sample Size")
abline(h=0.85, col=2, lwd=2, lty=2)
```

To sum up …

Two kinds of errors

	H_0 true	H_A true
Reject H_0 Specificity	Type I error	Good
Fail to reject H_0 Sensitivity	Good	Type II error

Type I error - false positive (“gullible”)

Type II error - false negative (“missed out on an opportunity”)

Goal: design a procedure that can ensure that

$$P(\text{Type I error}) \leq \alpha$$

yet still has small $P(\text{Type II error})$.

Significance level

By $P(\text{Type I error})$ we mean

$$P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

Think of significance level as our *level of gullibility*

- thinking I'm psychic when I'm random guessing
- declaring drug works when it actually makes no difference
- jury deciding "guilty" when person is innocent

Power

The **power** of a test is

$$P(\text{Reject } H_0 \mid H_A \text{ is true})$$

Power is test's **ability to detect** that alternative applies.

- detecting that drug works when it in fact does
- jury deciding "guilty" when person was guilty of crime

Note:

$$P(\text{Type II error}) = P(\text{Fail to reject } H_0 \mid H_A \text{ is true}) = 1 - \text{Power}$$

Rejection region approach

Rejection region approach

Classic Approach. Later comes the p-value approach

1. Specify H_0 and H_A
2. Determine **test statistic**

   Distribution to do calculations e.g. X~Binom(n, p)

   - figure out its sampling distribution under H_0
3. Determine **rejection region**

   This changes with p-value approach

   - specify for which observed values we will reject H_0
   - choose size of it to ensure significance level is α
4. **Decision**: Did observed value of test statistic fall in rejection region?
5. Check assumptions

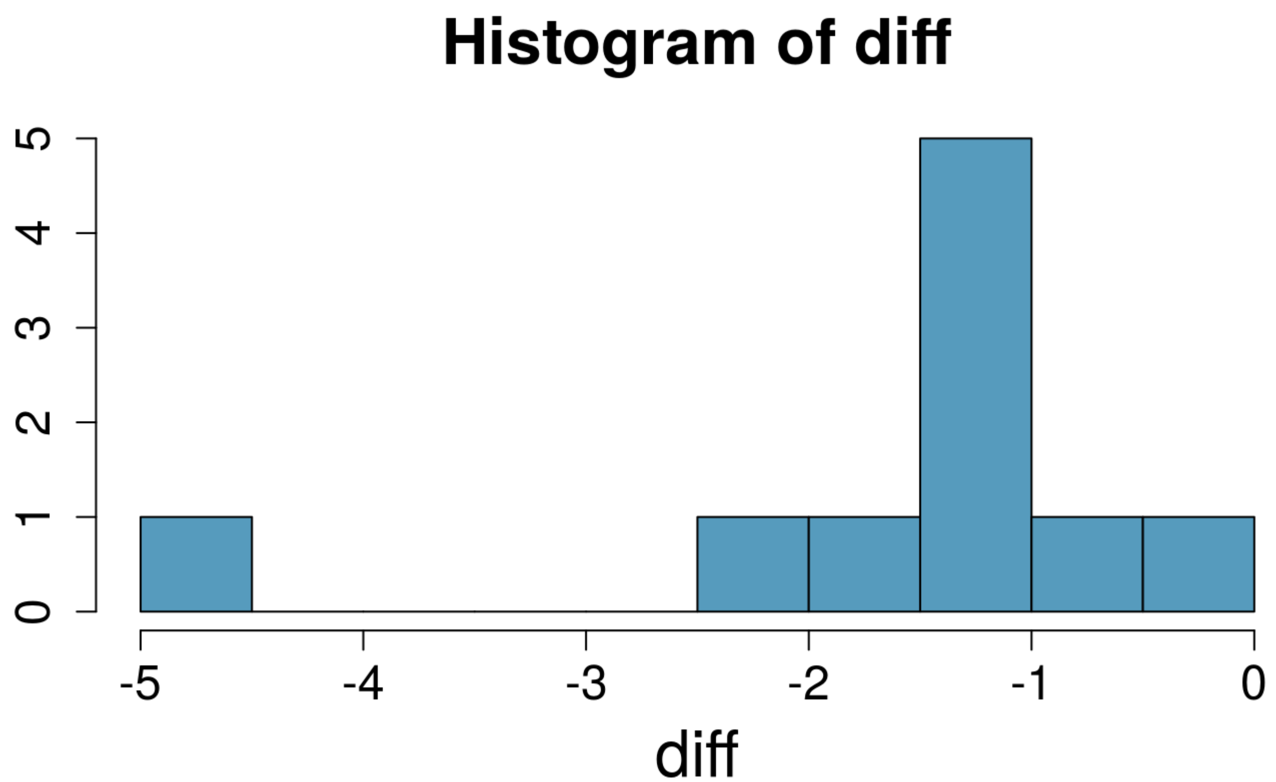
Example (from Gosset himself!)

Measure effect of sleeping drugs A and B on each of 10 people

	A	B	diff
person1	0.7	1.9	-1.2
person2	-1.6	0.8	-2.4
person3	-0.2	1.1	-1.3
person4	-1.2	0.1	-1.3
person5	-0.1	-0.1	0.0
person6	3.4	4.4	-1.0
person7	3.7	5.5	-1.8
person8	0.8	1.6	-0.8
person9	0.0	4.6	-4.6
person10	2.0	3.4	-1.4

Example (from Gosset himself!)

```
hist(diff, breaks=10)
```



Step 1: Identify hypotheses

Null hypothesis - “no difference between drugs”

- a hypothesis is a statement about the population parameter
- let μ be the (population) mean difference between drug A and drug B.
- $H_0 : \mu = 0$

Alternative hypothesis - “there is a difference between the drugs”

- $H_A : \mu \neq 0$
- this is called a “two-sided” hypothesis
- two-sided hypothesis should be your default choice

Step 2: Test statistic

$X_i =$ the difference in effect between the drugs for person i.

Test statistic - let's make decision based on $\bar{X}_n = \frac{X_1+\cdots+X_n}{n}$

Sample Statistic

Assuming X_i is approximately $N(\mu, \sigma)$, then

$$\bar{X}_n \approx N(\mu, \sigma/\sqrt{n}).$$

Standard Error

Under $H_0 : \mu = 0$, we have $\bar{X}_n \approx N(0, \sigma/\sqrt{n})$ or

$$\frac{\bar{X}_n - 0}{\sigma/\sqrt{n}} \approx N(0, 1)$$ Test Statistic

If σ were known we'd have a test statistic with known sampling distribution under the null!

Step 3: Rejection region

Rejection region - range of values of test statistic for which we will reject H_0 in favor of H_A.

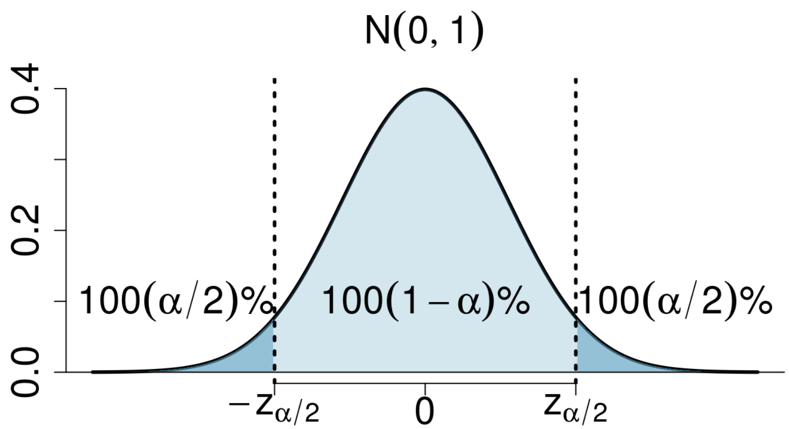
- look at H_A to decide whether low values, high values, or both would be considered evidence against H_0 in favor of H_A.

In example:

$H_A : \mu \neq 0$. So will reject if computed value of our test statistic $\frac{\bar{x}_n - 0}{\sigma/\sqrt{n}}$ is too high or too low.

What does "too high" or "too low" mean?

What does "too high" or "too low" mean?



$$\frac{\bar{X}_n - 0}{\sigma/\sqrt{n}} \approx N(0,1)$$

so if we calculate $\frac{\bar{x}_n - 0}{\sigma/\sqrt{n}}$ and it falls in tails, we'd reject H_0 in favor of H_A.

Step 3: Rejection region

Reject H_0 in favor of H_A if

$$\left| \frac{\bar{x}_n - 0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}.$$

doing so ensures Type I error rate is α:

$$P\left(\text{Reject } H_0 \,\middle|\, H_0 \text{ true}\right) = P\left(\left| \frac{\bar{X}_n - 0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \,\middle|\, \mu = 0\right)$$

$$= P\left(|N(0,1)| > z_{\alpha/2}\right) = \alpha$$

Step 4: Decision

Compute our test statistic (assume we know $\sigma = 1$):

```
alpha = 0.05; mu0 = 0; sigma = 1
xbar = mean(diff); n = length(diff)
(xbar - mu0) / (sigma / sqrt(n))
```

```
## [1] -4.996399
```

```
zvalue = -qnorm(alpha / 2)
zvalue
```

```
## [1] 1.959964
```

We reject H_0 in favor of H_A since $4.996 \geq 1.96$.

Step 5: Check assumptions

Assumptions made:

- independence of X_i's
- approximate normality of X_i's

In reality, we don't know $\sigma = 1$. How does the above change?

Step 1: Identify hypotheses

(Unchanged)

- $H_0 : \mu = 0$
- $H_A : \mu \neq 0$

Step 2: Test statistic

X_i = the difference in effect between the drugs for person i.

Test statistic - let's make decision based on $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$

Assuming X_i is approximately $N(\mu, \sigma)$, then

$$\bar{X}_n \approx N(\mu, \sigma/\sqrt{n}).$$

Under $H_0 : \mu = 0$, we have $\bar{X}_n \approx N(0, \sigma/\sqrt{n})$ or

$$\frac{\bar{X}_n - 0}{\sigma/\sqrt{n}} \approx N(0, 1)$$

However, since σ is unknown, we can't calculate its value and so it is **not a usable test statistic.**

Step 2: Test statistic

Problem: we don't know σ, so we can't compute

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

What did we do for confidence intervals?

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

where

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

Flashback Why is it not normal?

Intuitively,

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \qquad \text{(let's call this } T_n\text{)}$$

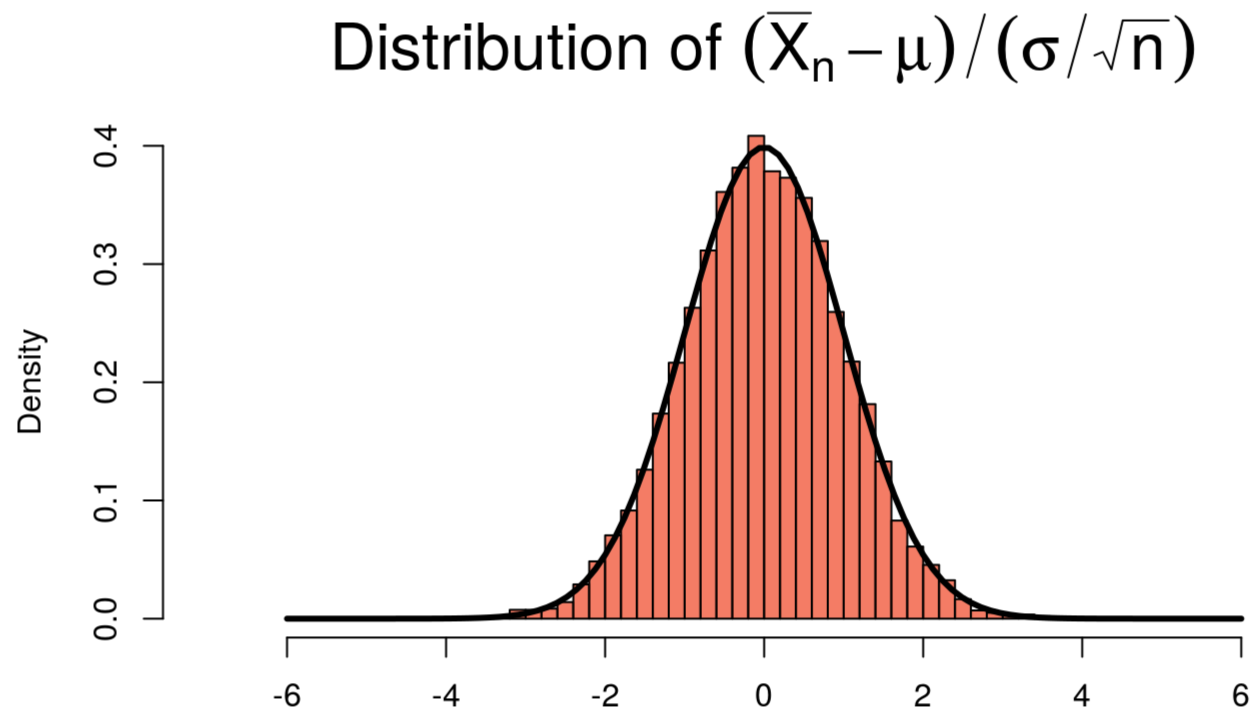
has more variability “in it” than

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

since S_n is also random.

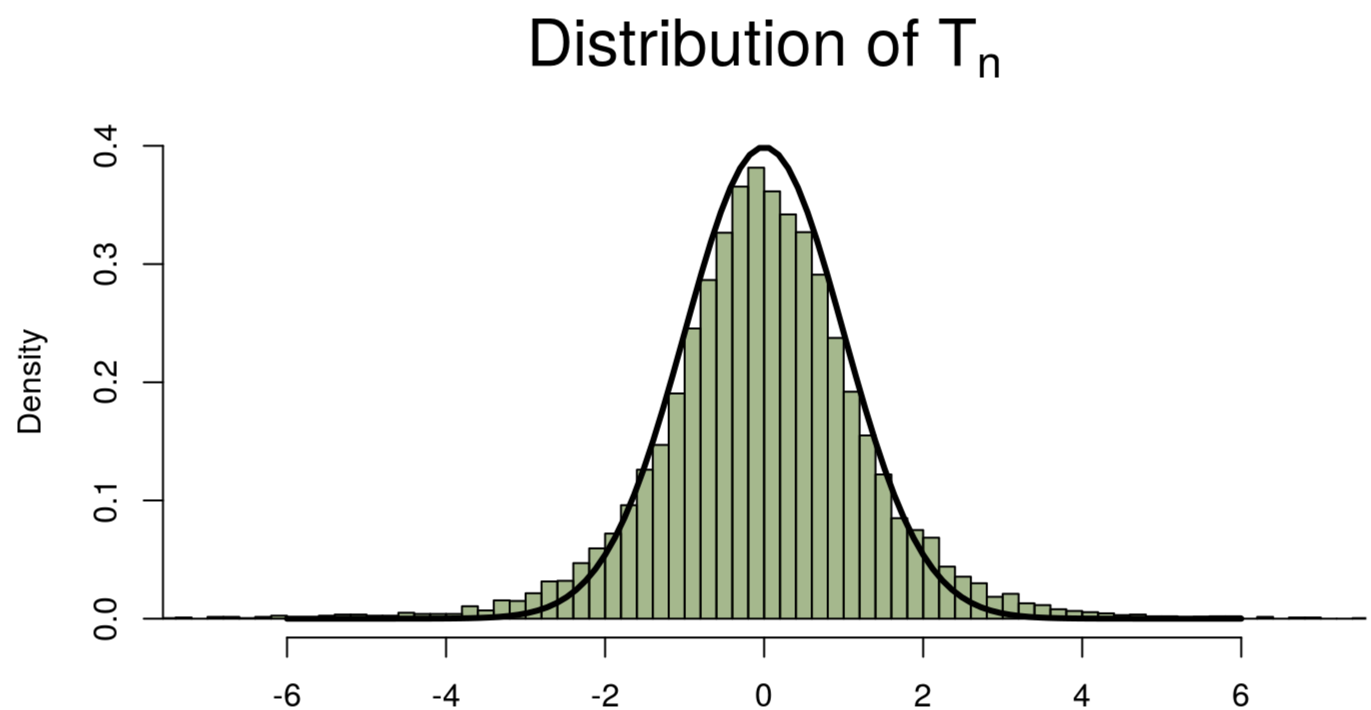
Monte Carlo simulation (normal data, n=5)

Draw $X_1, \ldots, X_5 \sim N(\mu, \sigma)$: See $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$



Monte Carlo simulation (normal data, n=5)

Same as before. See $T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ has heavier tails than $N(0,1)$



Student's t-distribution

If $X_1, \ldots, X_n \sim N(\mu, \sigma)$ are independent, then

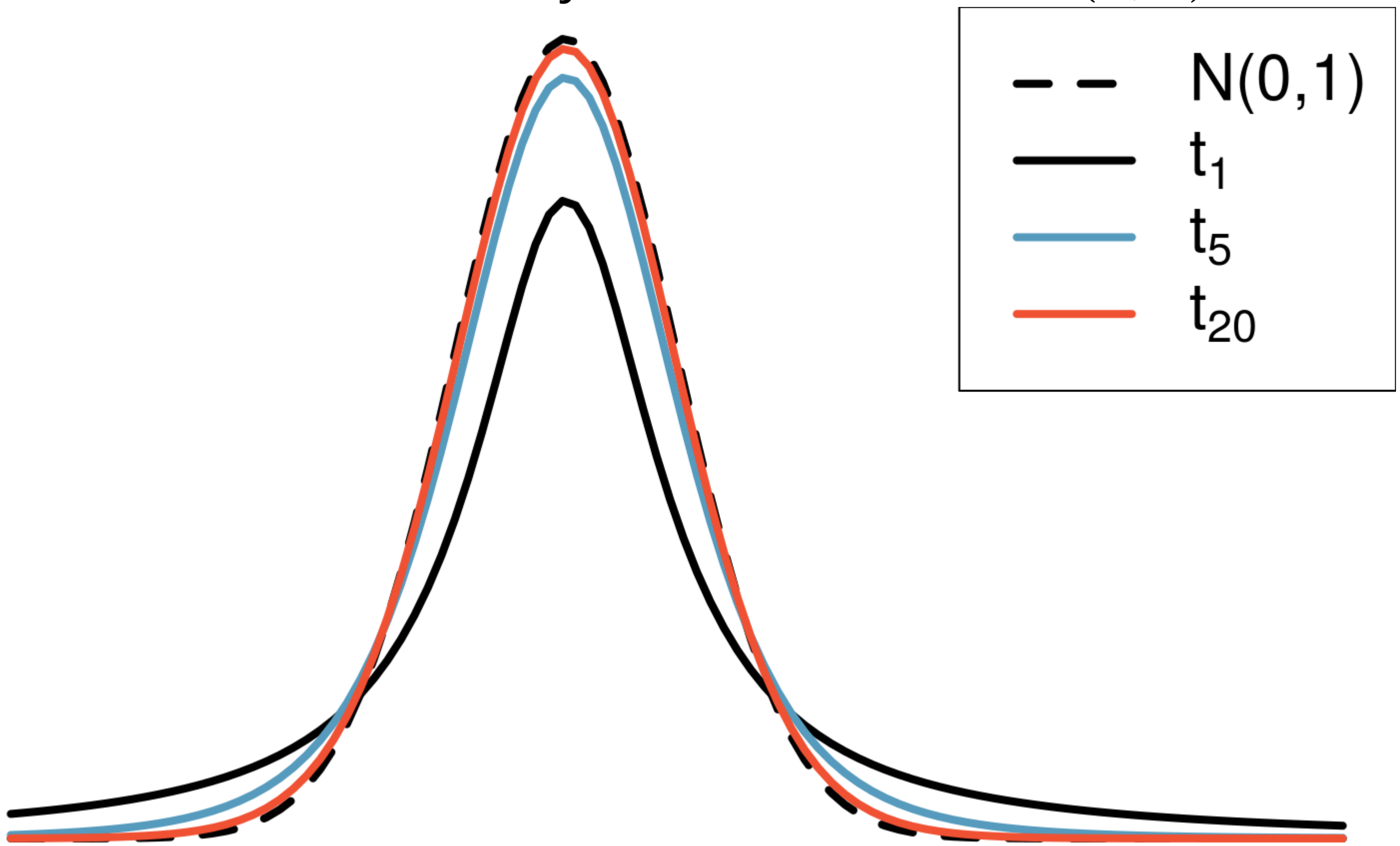
$$T_n = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t_{n-1}$$

In words, we say that T_n has a **t-distribution with $n-1$ degrees of freedom**.

t_{n-1} denotes this distribution.

Student's t-distribution

For small n has noticeably heavier tails than $N(0, 1)$.



Flashforward Step 2: Test statistic

Let's use

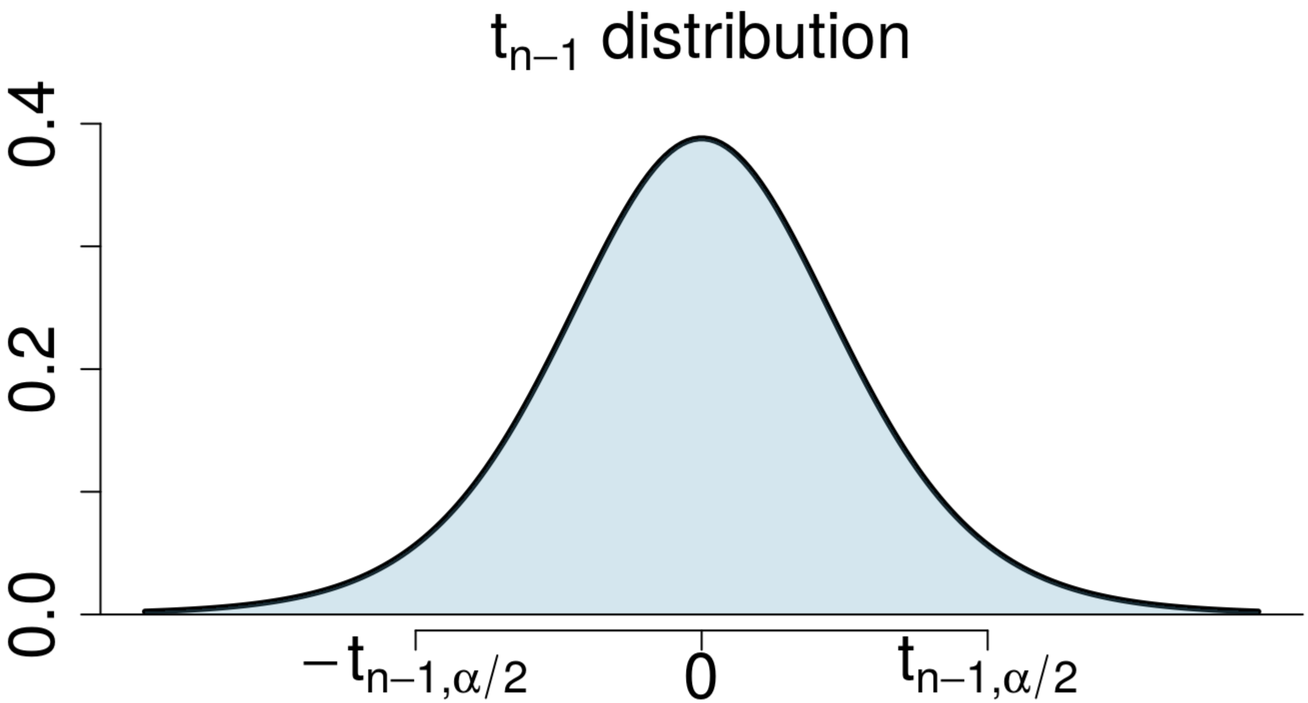
$$\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

as our test statistic.

Under $H_0 : \mu = \mu_0$, its sampling distribution is known: t_{n-1}

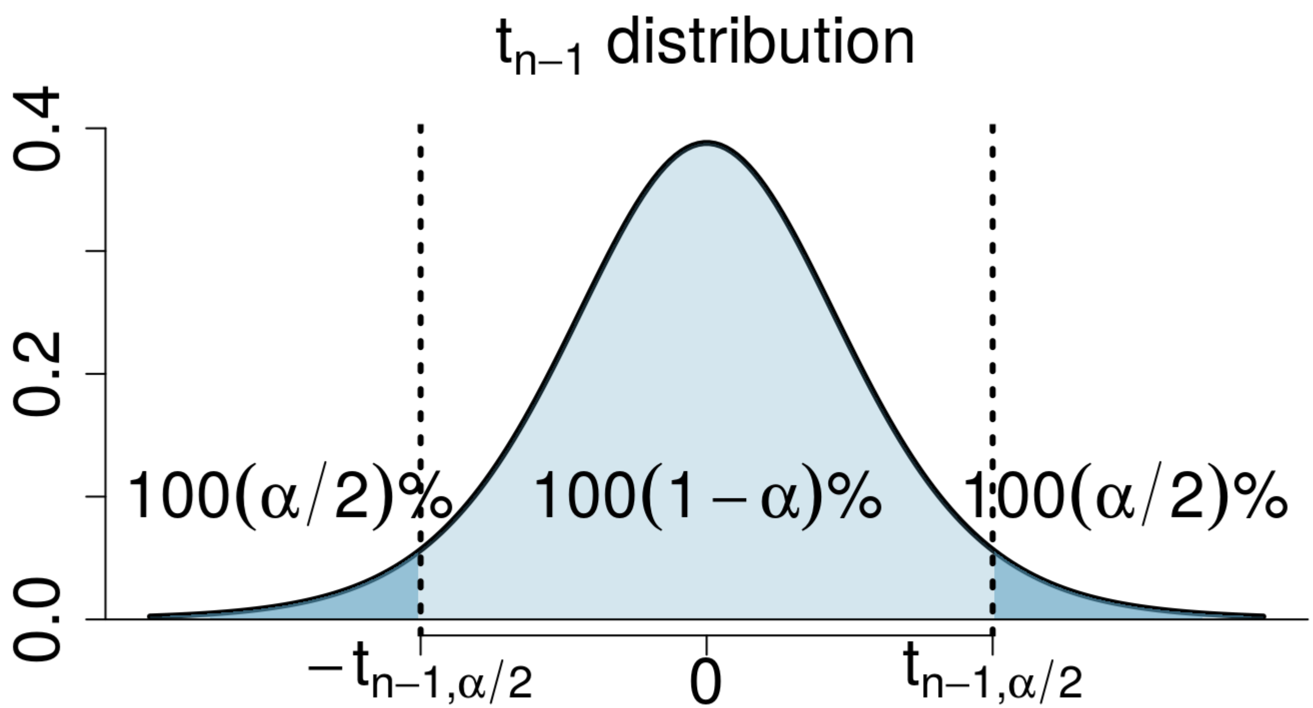
Step 2: Test statistic

Distribution of $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ under H_0



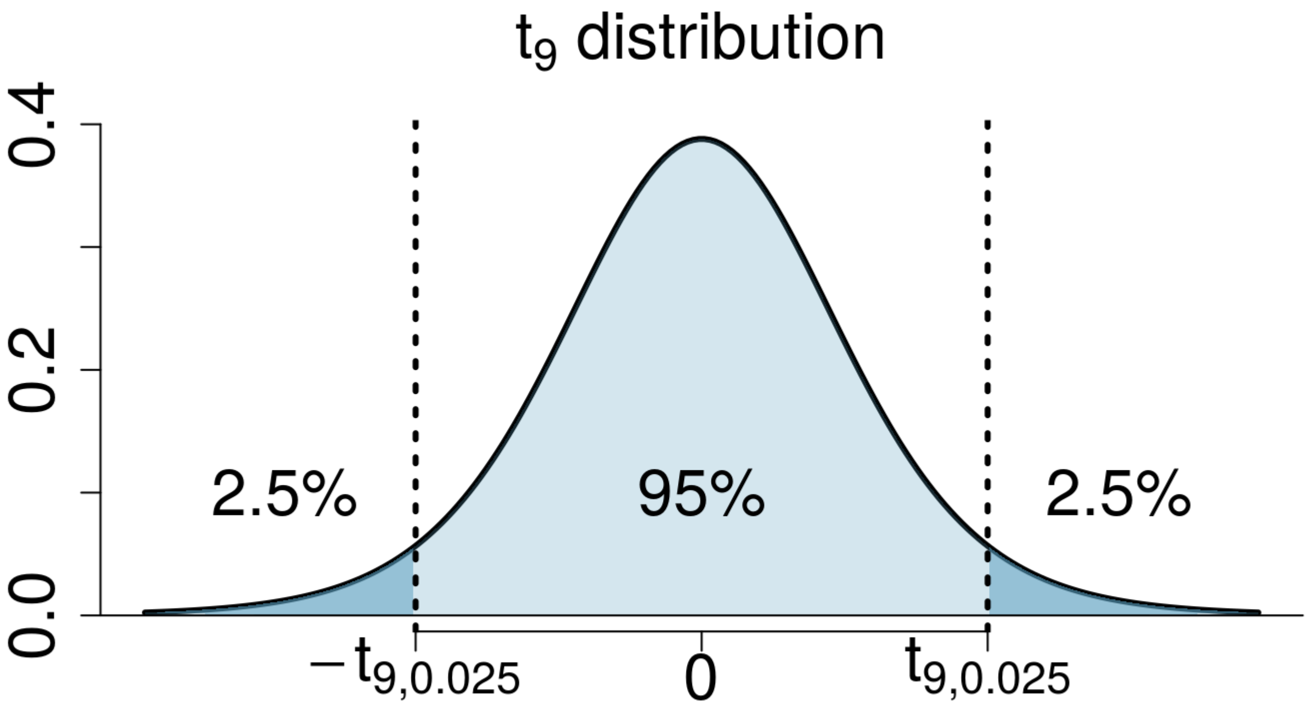
Step 3: Rejection region

Distribution of $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ under H_0

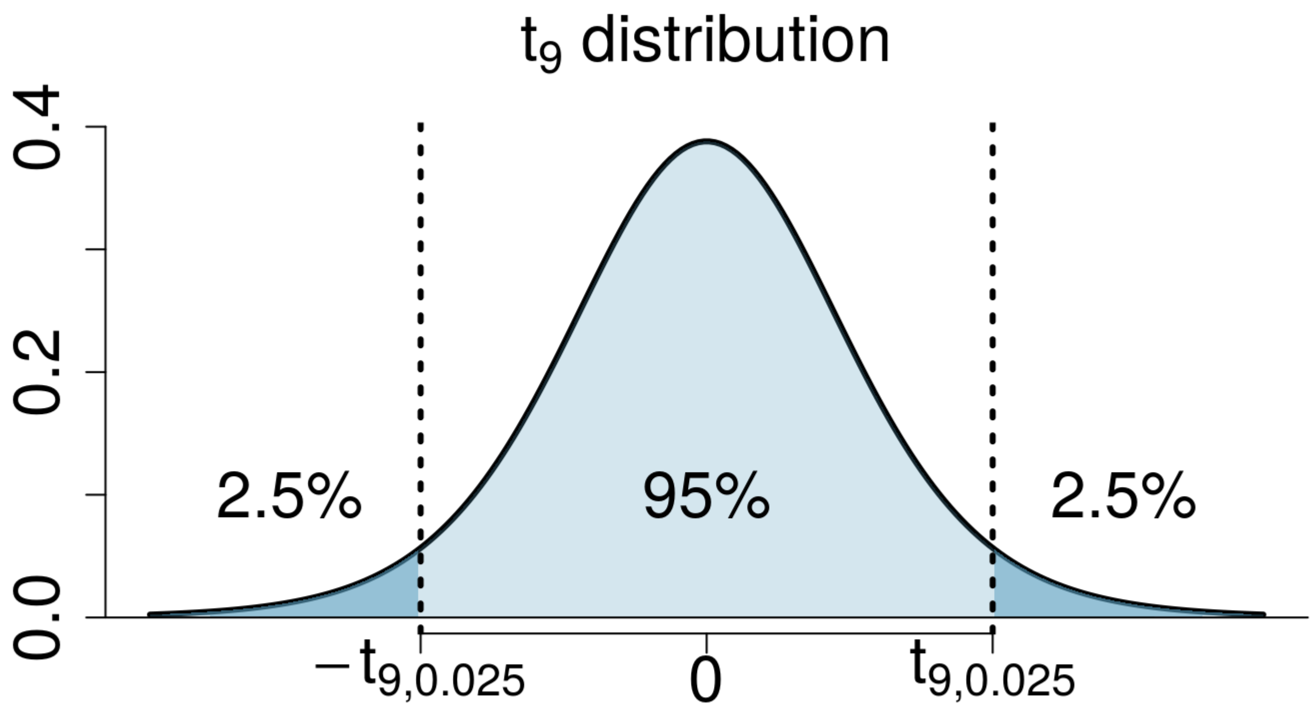


Step 3: Rejection region

Sleep example: $\frac{\bar{X}_n - 0}{S_n/\sqrt{n}}$ under H_0 has distribution t_9



Step 3: Rejection region



Reject H_0 in favor of H_A if computed $\left|\frac{\bar{x}_n - 0}{s/\sqrt{n}}\right| > t_{9,0.025}$.

Step 4: Decision

Compute our test statistic:

```
alpha = 0.05; mu0 = 0
xbar = mean(diff); n = length(diff); s = sd(diff)
(xbar - mu0) / (s / sqrt(n))
```

```
## [1] -4.062128
```

```
tvalue = -qt(alpha / 2, df = n - 1)
tvalue
```

```
## [1] 2.262157
```

We reject H_0 in favor of H_A since $4.062 \geq 2.262$.

Note

For t-test, our cutoff is $t_{n-1,\alpha/2}$ which is bigger than $z_{\alpha/2}$.

$$z_{0.025} = 1.960$$

$$t_{9,0.025} = 2.262$$

Intuition for higher cutoff: We're less surprised by a large value of test statistic when if we were using S_n.

Step 5: Check assumptions

t-test assumes X_i's are normal. How to evaulate?

- matters most when n is small
- hardest to verify when n is small (unfortunate!)

p-value approach

p-value approach

1. Specify H_0 and H_A
2. Determine **test statistic**
   - figure out its sampling distribution under H_0
3. Compute the **p-value** based on the particular sample of data you collected
4. **Decision**: Did p-value fall below α? If so, reject H_0 in favor of H_A.
5. Check assumptions

Steps 1 and 2

Identical to what was done in rejection region approach

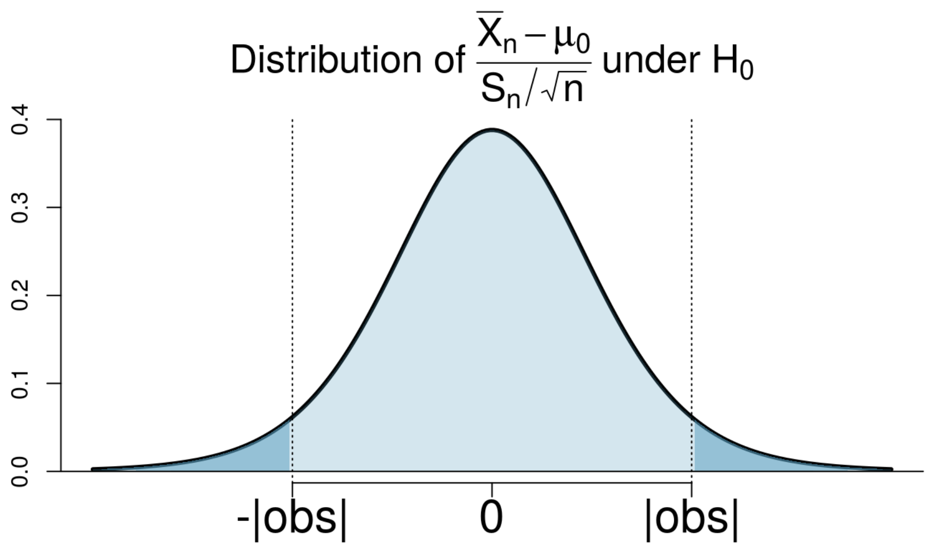
Step 3: Compute the p-value

> The **p-value** is the probability under the null hypothesis of seeing something as extreme or more extreme than what was actually observed.

- "**extreme**" is defined by H_A.
- in example, $H_A : \mu \neq 0$, so "extremeness" means how far $\bar{X}_n$ is from 0, that is $|\bar{X}_n - 0|$.
- "**probability under the null**" - we know $\frac{\bar{X}_n - 0}{S_n/\sqrt{n}} \sim t_{n-1}$
- "**actually observed**" - we observed $\frac{\bar{x}_n - 0}{s/\sqrt{n}} = -4.06$

Putting it all together: what is the probability that a t_{n-1} random variable would be larger than 4.06 or smaller than -4.06?

Always draw a picture!



- sampling distribution is a t_{n-1} distribution
- obs is the observed value (realization of test statistic):

$$\frac{\bar{x}_n - \mu_0}{s/\sqrt{n}}$$

Calculating p-value

Want

$$P(t_{n-1} < -|obs|) + P(t_{n-1} > |obs|) = 2P(t_{n-1} < -|obs|)$$

In example:

```
mu0 = 0 # null value
xbar = mean(diff); s = sd(diff); n = 10
obs = (xbar - mu0) / (s / sqrt(n))
pvalue = 2*pt(-abs(obs), df = n - 1)
pvalue
```

```
## [1] 0.00283289
```

Step 4: Decide

- If p-value is less than α, we reject H_0 in favor of H_A.
- this step is optional if you've calculated the p-value. You can leave it up to the reader whether this is "statistically significant"

Step 5: Check assumptions

Same as before.

Other remarks

What makes a p-value small?

Recall for two-sided t-test, our p-value was

$$2P\left(t_{n-1} > \left|\frac{\bar{x}_n - \mu_0}{s/\sqrt{n}}\right|\right)$$

where $\bar{x}_n$ and s are the computed values from your data.

- p-value is small when $\left|\frac{\bar{x}_n - \mu_0}{s/\sqrt{n}}\right|$ is large.
- this occurs when:
  - $|\bar{x}_n - \mu_0|$ is large (happens when true μ is far from μ_0)
  - s is small (happens when σ is small)
  - n is large (in your control!)

What makes a p-value small?

Consequence:

- As long as true μ is not exactly equal to μ_0, we can get small p-values by increasing n.
- *cynical view:* p-values just reflect your amount of effort (sample size) relative to what's there,

$$\frac{|\mu - \mu_0|}{\sigma}.$$

Practical significance

Important distinction

> Statistical significance $\neq$ Practical significance

In example: If drug A increases sleep by 1.2 minutes over drug B, do we care?

- relevant question since we could still get a p-value < 0.0001 with a large enough n.
- essential to consider the **effect size**
  - such as $|\mu - \mu_0|$
  - or standardized $|\mu - \mu_0|/\sigma$

Duality between testing and confidence intervals

There's a connection!

1. Recall **level α test** for $H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq \mu_0$:

> Reject H_0 when
>
> $$\frac{|\bar{x}_n - \mu_0|}{s_n/\sqrt{n}} > t_{n-1,\alpha/2}$$

1. **$100(1-\alpha)\%$ confidence interval** for μ:

> $$\left[\bar{x}_n - t_{n-1,\alpha/2} s_n/\sqrt{n}, \bar{x}_n + t_{n-1,\alpha/2} s_n/\sqrt{n}\right]$$

Observe: μ_0 in confidence interval is equivalent to failing to reject H_0!

There's a connection!

Justification: μ_0 being in CI means

$$\bar{x}_n - t_{n-1,\alpha/2} s_n/\sqrt{n} \leq \mu_0 \leq \bar{x}_n + t_{n-1,\alpha/2} s_n/\sqrt{n}$$

or

$$|\bar{x}_n - \mu_0| \leq t_{n-1,\alpha/2} s_n/\sqrt{n}$$

or

$$\frac{|\bar{x}_n - \mu_0|}{s_n/\sqrt{n}} \leq t_{n-1,\alpha/2}$$

Compare to

> Reject H_0 when
>
> $$\frac{|\bar{x}_n - \mu_0|}{s_n/\sqrt{n}} > t_{n-1,\alpha/2}$$

New interpretation of CIs

A $100(1-\alpha)\%$ confidence interval consists of all the values of μ_0 for which a test of the form

$$
\begin{aligned}
H_0 &: \mu = \mu_0 \\
H_A &: \mu \neq \mu_0
\end{aligned}
$$

would fail to reject H_0 at the α significance level.

Why this is useful

Suppose we are interested in testing

$$H_0 : \mu = 0$$
$$H_A : \mu \neq 0.$$

Suppose we get a 95% confidence interval: [-0.6, 1.1]

Test would fail to reject at 0.05 significance level because the interval [-0.6, 1.1] includes 0.

Sanity check: what happens to both if you increase α?