Hypothesis Tests

Jim Booth

October 17, 2019

Recap: Inference

Sampling distribution: A probabilistic description of how the observed values of a numerical summary statistic (e.g., sample mean) behave under repeated SRS.

This concept underlies all basic statistical inference procedures – its importance cannot be overstated!

In practice: we only collect one sample.

Question: how can we combine the information from a single SRS about a population parameter with our knowledge of sampling distributions in order to perform statistical inference?

Central Limit Theorem

If $X_1, \ldots, X_n$ are independent draws from a distribution with mean μ and standard deviation σ, then for large n, the sample mean $\bar{X}_n$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$:

> $$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- remarkable since individual X_i 's don't have to look at all like a normal distribution
- how large should n be? Depends, but if distribution of X_i 's is not strongly skewed, say $n \geq 30$

Two primary goals

1. A **confidence interval (CI)** - a range of plausible values for a (population) parameter, based on the data obtained from our observed sample.

2. A **hypothesis (or significance) test** - an assessment of whether the observed value of a statistic computed using the sample data is consistent with or divergent from some hypothesized value of the (population) parameter.

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

Each picks a number between 1 and 4

I "guess" each person's number

x = Number that I get 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

- Note that for a binomial RV

$$P(X \geq 25) = 1 - P(X \leq 24)$$

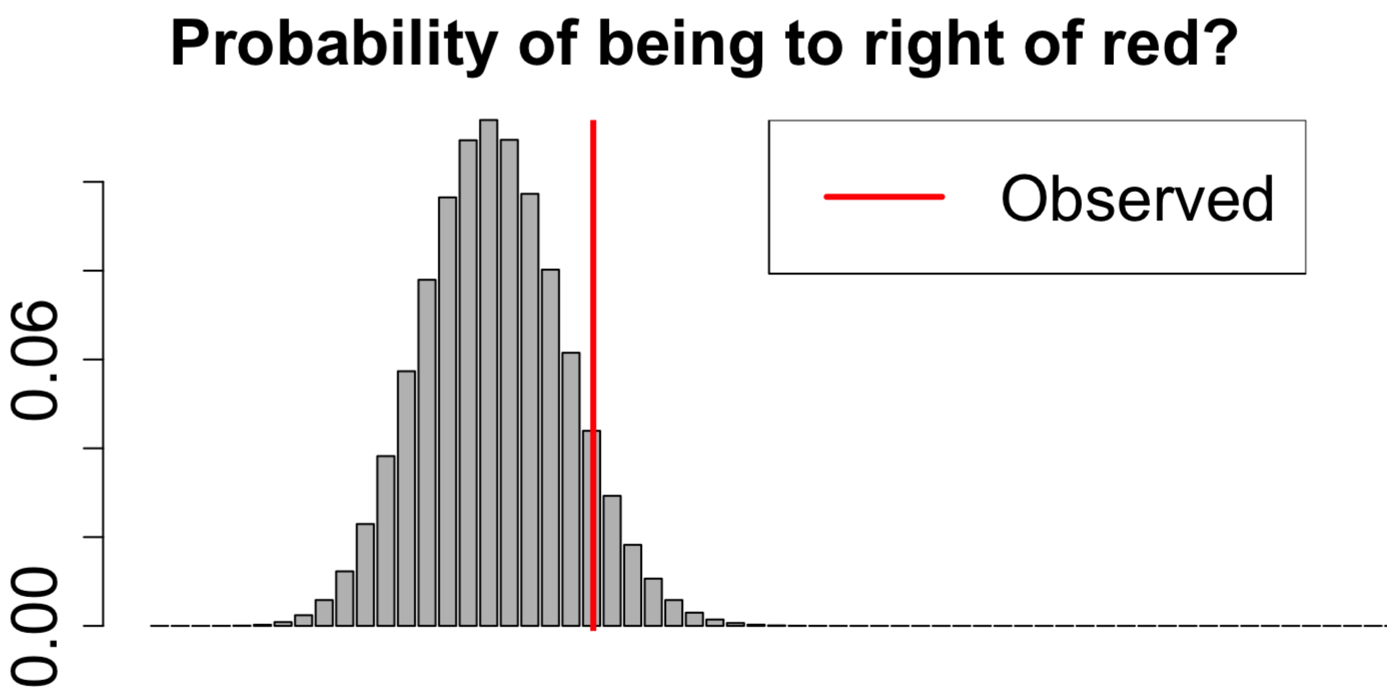
```
n=65 # number of trials
x=25 # number of successes
p = 1/4 # calculate under null hypothesis
1 - pbinom(x-1, size = n, prob = p)
```

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

- 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



$$P(X \geq 25) = 1 - P(X \leq 24)$$

How small is small enough?

- 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
- Suppose 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
Fail to reject H_0	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 randomly 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:

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

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

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

with, say, $\alpha = 0.05$.

Finding cutoff

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

```r
alpha = 0.05
n = 65
p = 1/4 # under null
quantile = qbinom(1-alpha, n, p) # this is c - 1
cutoff = quantile + 1
cutoff
```

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

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

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

Rejection region

Our level $\alpha \leq 0.05$ 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 $X \sim \text{Binomial}(n, p)$.

Power

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

```
cutoff
```

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

```
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

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

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

```
cutoff
```

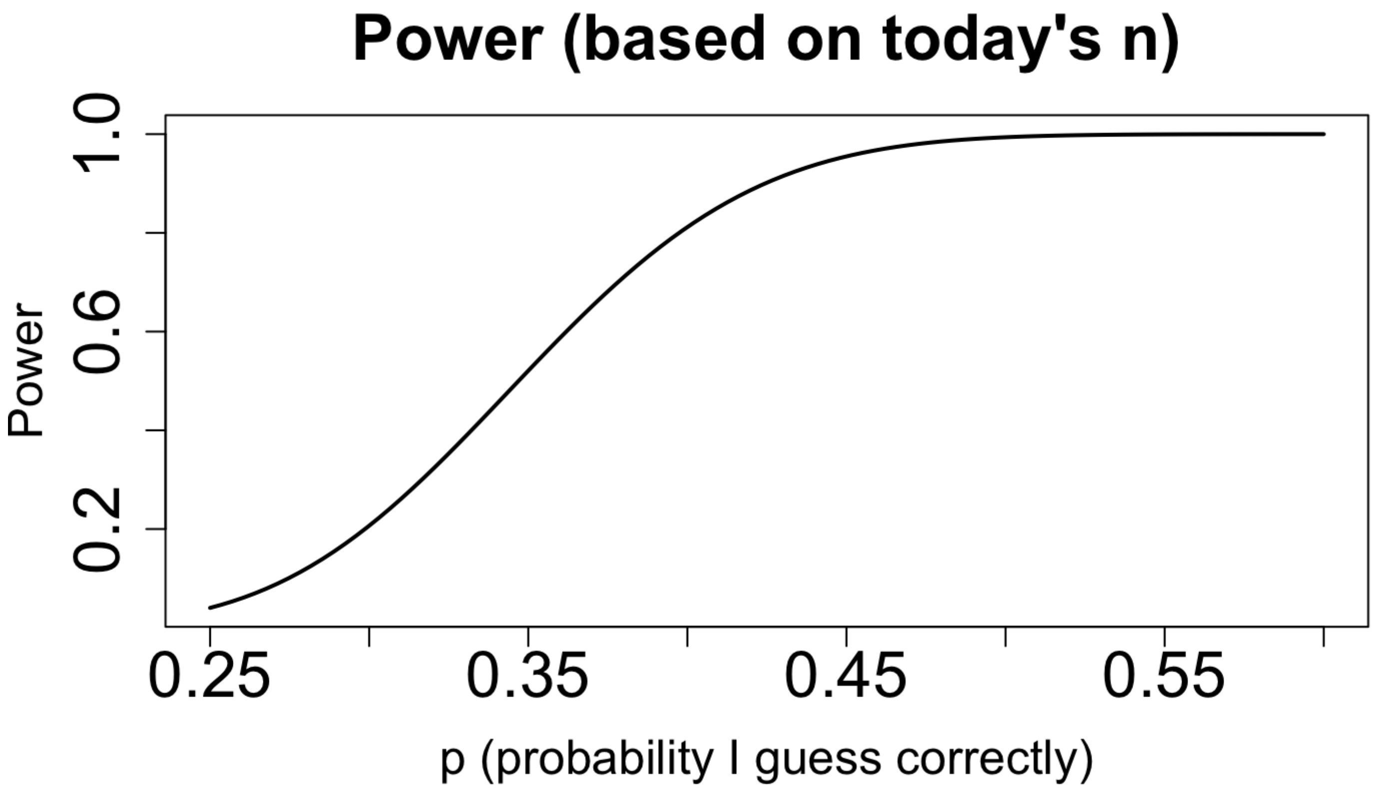
```
## [1] 23
```

```
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?

```r
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
```

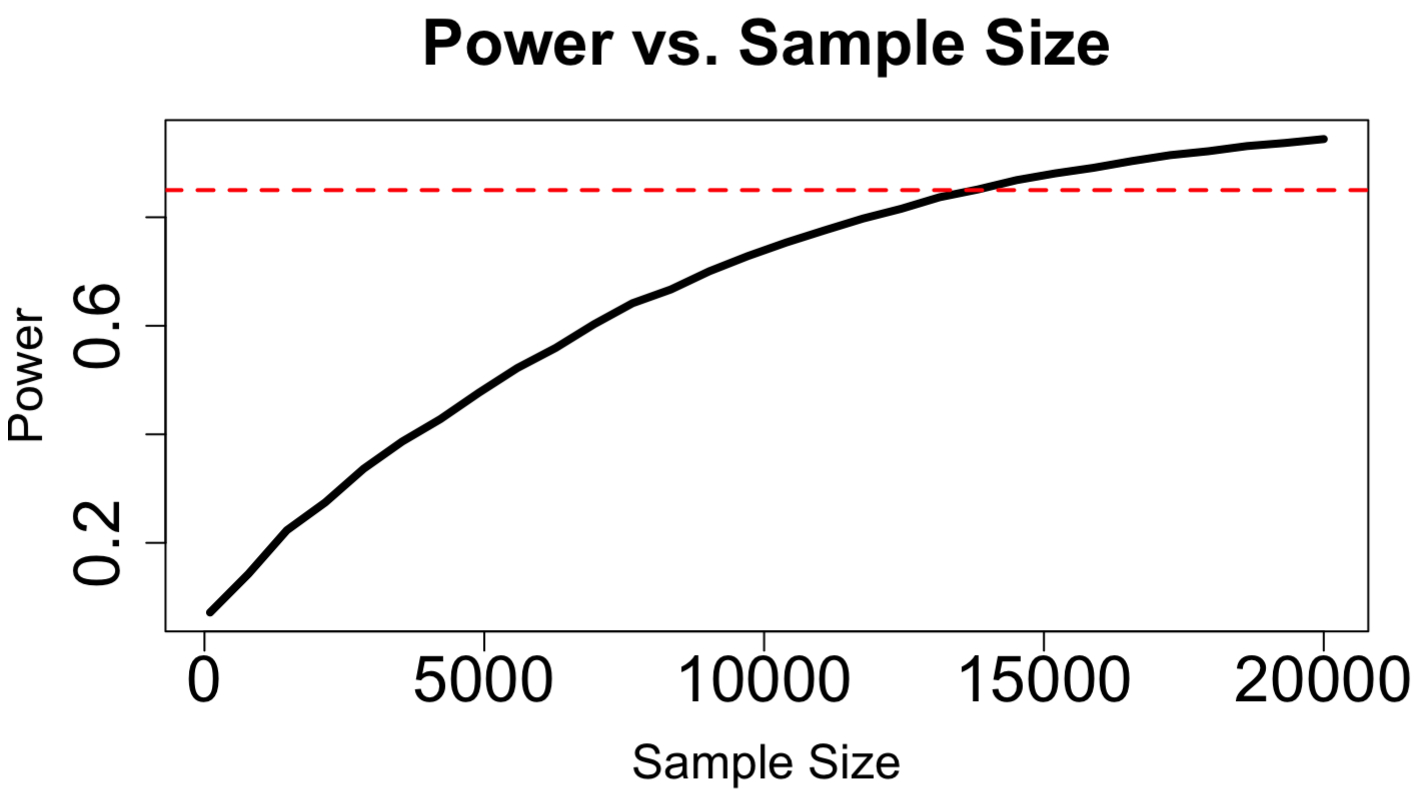
A power calculation

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)
}

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	Type I error	Good
Fail to reject H_0	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

1. Specify H_0 and H_A
2. Determine **test statistic**
   - figure out its sampling distribution under H_0
3. Determine **rejection region**
   - 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

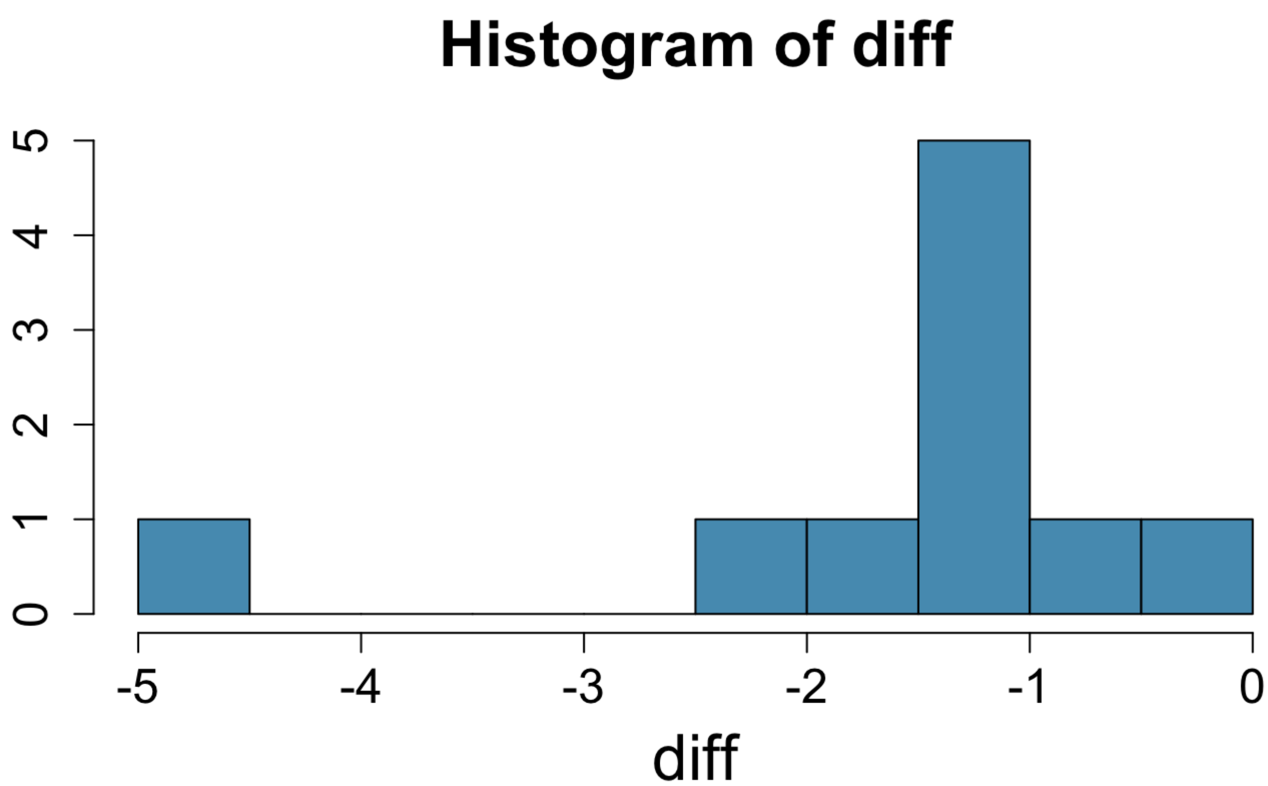
Example (from Gosset himself!)

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

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 $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$

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

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

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.

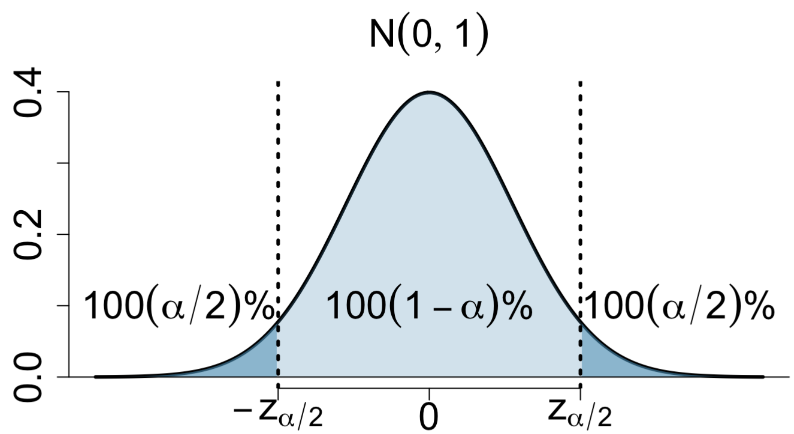
- look at H_A to decide whether low values, high values, or both would be considered evidence against H_0 in favor of H_A.

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(1-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 σ. How does the above change?