

Confidence intervals for  
the population variance  $\sigma^2$   
based on the sample variance  $s^2$

# Confidence interval for the population variance $\sigma^2$

- Up until now we were calculating the confidence interval on the **population average  $\mu$**
- What if one wants to put **confidence interval on the population variance  $\sigma^2$** ?
- We know an unbiased estimator of  $\sigma^2$ :

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

- How to determine the confidence interval?



$$\vec{x} = (x_1, x_2, \dots, x_n)$$

$$x_i \rightarrow x_i - \bar{x}$$

$$y = |\vec{x}|^2 = \sum x_i^2 = (n-1)s^2$$

$$\sum_{i=1}^n x_i = 0$$

$$P(\vec{x}) d|\vec{x}| \sim \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) dx_i$$

(left the last one since  $x_n = -\sum_{i=1}^{n-1} x_i$ )

$$|\vec{x}| = \sqrt{y}$$

sphere  
area  $\sim$   
 $|\vec{x}|^{n-2}$

$$d|\vec{x}| = \frac{1}{\sqrt{y}} dy$$

$$\prod dx_i \sim |\vec{x}|^{n-2} d|\vec{x}|$$



$$P(y) dy = y^{\frac{n-1}{2}-1} \exp\left(-\frac{y}{2}\right) dy$$

# 8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

---

## Definition

(Eq. 8-17)

Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and let  $S^2$  be the sample variance. Then the random variable

$$\chi^2 = \frac{(n - 1) S^2}{\sigma^2} \quad (8-17)$$

has a chi-square ( $\chi^2$ ) distribution with  $n - 1$  degrees of freedom.

# 8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$$X = (n-1)S^2 / \sigma^2$$

We know  $n, S^2$

want to estimate  $\sigma^2$

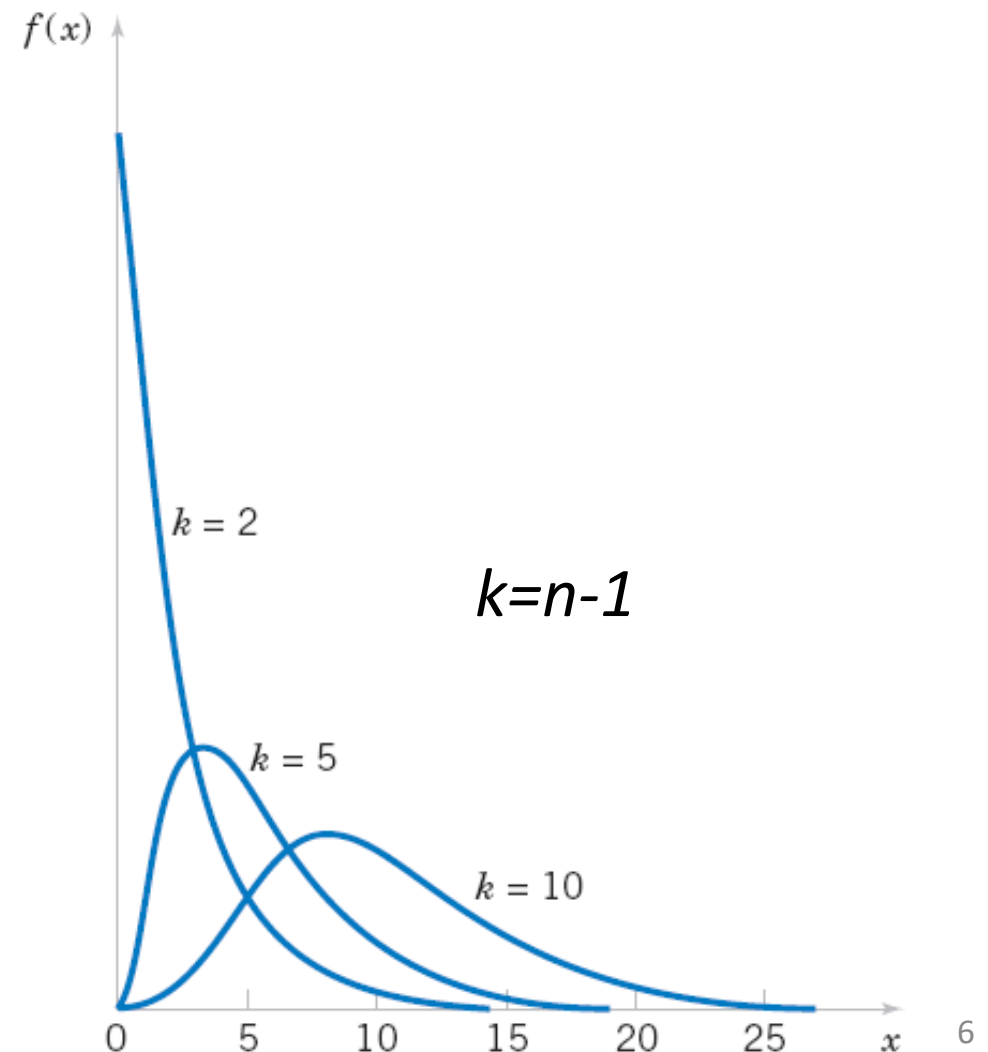
$$f(x, n) \sim x^{(n-1)/2-1} \exp(-x/2)$$

It is just Gamma PDF  
with  $r = (n-1)/2$ , and  $\lambda = 1/2$

Mean value:  
 $n-1$

Standard deviation:

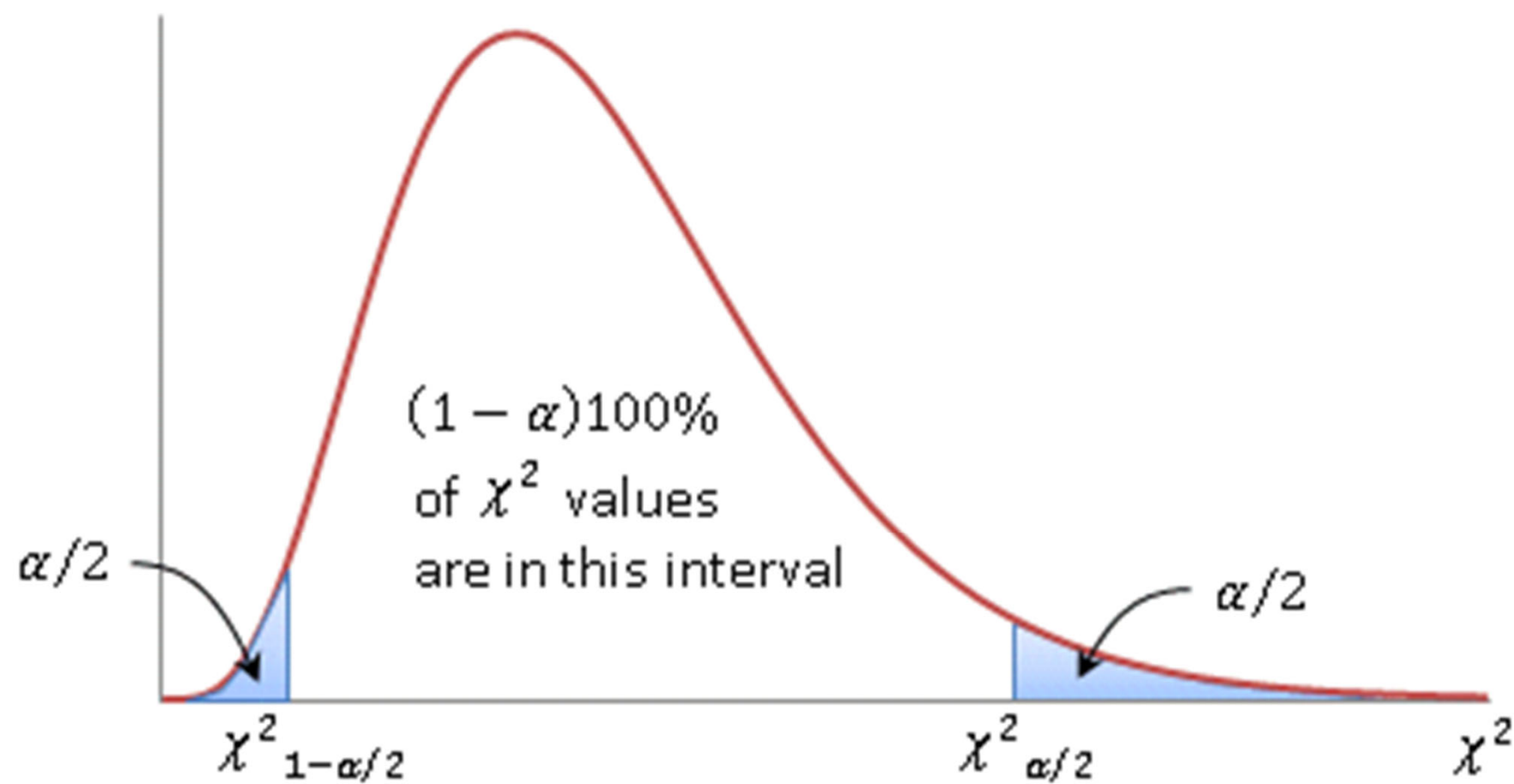
$$\sqrt{2(n-1)}$$



# Play with Mathematica notebook

<http://demonstrations.wolfram.com/ChiSquaredDistributionAndTheCentralLimitTheorem/>

By Peter Falloon



$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$



# 8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

---

## Definition

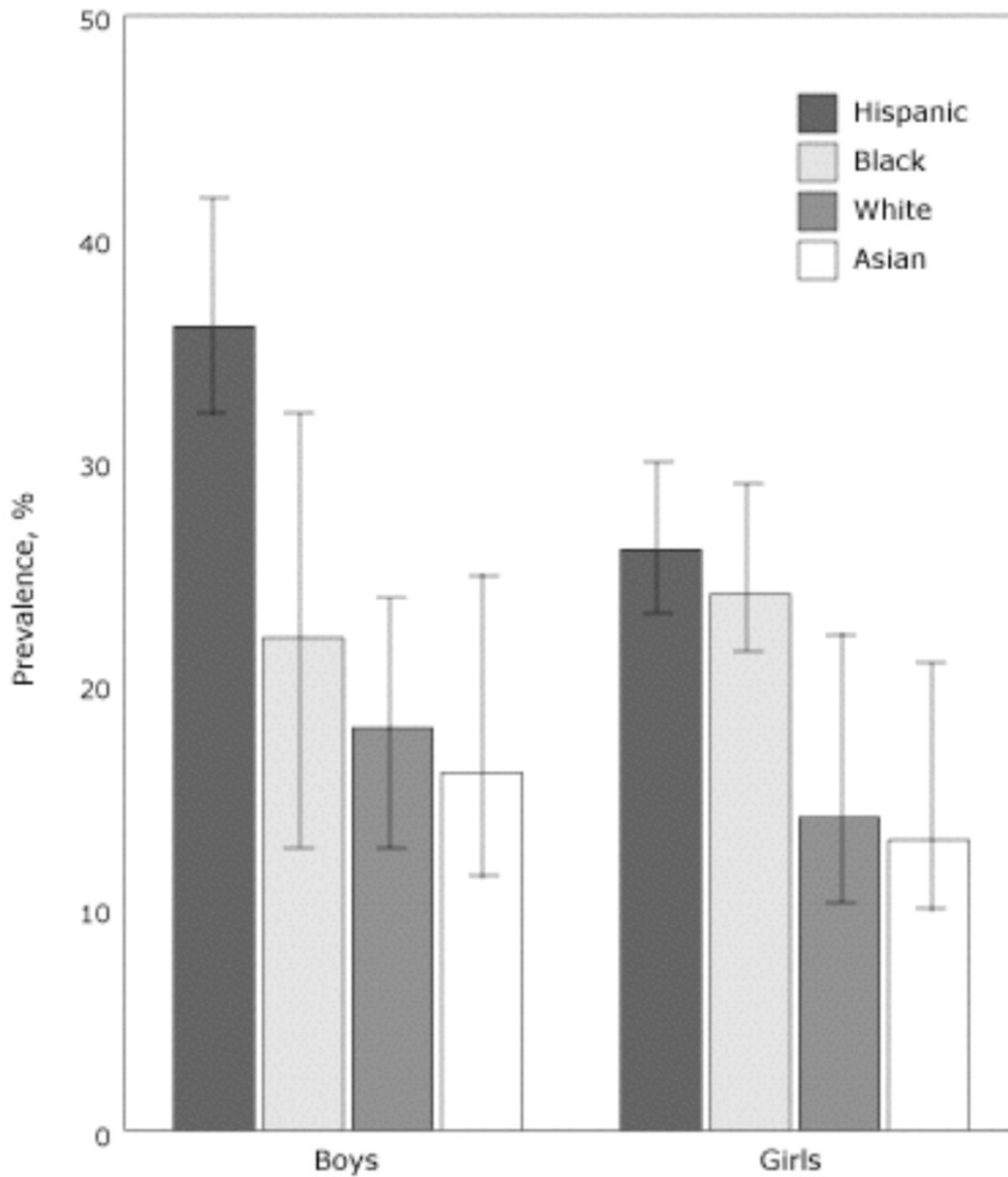
(Eq. 8-19)

If  $s^2$  is the sample variance from a random sample of  $n$  observations from a normal distribution with unknown variance  $\sigma^2$ , then a **100(1 -  $\alpha$ )% confidence interval on  $\sigma^2$**  is

$$\frac{(n - 1)s^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n - 1)s^2}{\chi_{1-\alpha/2, n-1}^2} \quad (8-19)$$

where  $\chi_{\alpha/2, n-1}^2$  and  $\chi_{1-\alpha/2, n-1}^2$  are the upper and lower 100 $\alpha$ /2 percentage points of the chi-square distribution with  $n - 1$  degrees of freedom, respectively. A **confidence interval for  $\sigma$**  has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.

# Confidence estimates of the population proportion



**Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.**

**(source: CDC.GOV)**

Collect a sample of BMI values  
 Obese means  $BMI > 30$

**What do those bars actually mean?**



# Large sample confidence estimate of population proportion

- Want to know the **fraction  $p$  of the population** that belongs to a class, e.g., the class “obese” kids defined by BMI>30.
- Each variable is a Bernoulli trial with one parameter  $p$ . We can use **moments** or **MLE estimator** to estimate  $p$
- Both give the same estimate: **sample fraction  $\hat{p}=(\# \text{ of obese kids in the sample})/(\text{sample size } n)$**
- How to put confidence bounds on  $p$  based on  $\hat{p}$
- # of obese kids in the sample follows the binomial distribution: “success” = sampled kid is obese : -(  
 $p$  – probability of success,  $1-p$  – failure
- Expected # of successes is  $np$  → Expected fraction of successes is  $p$
- Standard deviation of # of successes is  $\sqrt{np(1-p)}$  →  
Standard deviation of fraction of successes is  $\sqrt{p(1-p)/n}$

# 8-5 A Large-Sample Confidence Interval For a Population Proportion

---

## Normal Approximation for Binomial Proportion

If  $n$  is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

is approximately standard normal.

The quantity  $\sqrt{\hat{p}(1-\hat{p})/n}$  is the standard error of the point estimator  $\hat{p}$ .

## 8-5 A Large-Sample Confidence Interval For a Population Proportion (Eq. 8-23)

---

If  $\hat{p}$  is the proportion of observations in a random sample of size  $n$  that belongs to a class of interest, an approximate  $100(1 - \alpha)\%$  confidence interval on the proportion  $p$  of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (8-23)$$

where  $z_{\alpha/2}$  is the upper  $\alpha/2$  percentage point of the standard normal distribution.

This interval is known as the Wald interval (Wald and Wolfowitz, 1939).

Did you know that M&M's<sup>®</sup> Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

<http://www.scientificameriken.com/candy5.asp>

“To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (<http://us.mms.com/us/about/products/milkchocolate/>). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.

On average, our new mix of colors for M&M'S<sup>®</sup> Chocolate Candies is:

M&M'S<sup>®</sup> Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S<sup>®</sup> Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S<sup>®</sup> Kids MINIS<sup>®</sup>: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S<sup>®</sup> Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S<sup>®</sup> Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA  
A Division of Mars, Incorporated



How to estimate these probabilities from a finite sample and how to set confidence interval on these estimates?



Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI on the percentage of blue M&Ms to be less than +/- 4%  
Same question for red M&Ms?



Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?



How large is a sample needed for 95% CI on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

For blue M&Ms  $p = 0.24$

$$1.96 \sqrt{\frac{0.24(1-0.24)}{n}} < 0.04$$

$$n > \left(\frac{1.96}{0.04}\right)^2 0.24 \times (1-0.24) = 438 \text{ M\&Ms or}$$

~ 2 x 7oz bags with 210 candies each

For red M&Ms  $p = 0.13$

$$n > \left(\frac{1.96}{0.04}\right)^2 \times 0.13 \times (1-0.13) \approx 271 \text{ M\&Ms or}$$

~ 1 x 7oz bag

# Hypothesis testing: one sample

Is P53 gene expressed at a **lower level** in **cancer** patients than in **healthy** people?

- We are interested if a P53 gene expression is **lowered** in **population of cancer patients** compared to the **healthy population**.
- We know that mean gene expression in the **healthy population** is  $\mu_h = 50$  mRNAs/cell. We are interested in deciding whether or not the mean expression in **cancer population** is **lower than** in **healthy population**. Let's call hypothesis  $H_1$ . Here  $H_1$  is **one-sided**
- If we asked: cancer is **not equal** to healthy  $H_1$  would be a **two-sided hypothesis**
- Assume we have a sample of **100 cancer patients** with **sample mean  $\bar{x} = 48$  mRNAs/cell** and **standard deviation  $\sigma = 10$  mRNA/cell**
- Can we use our sample to reject the “business as usual” or **null hypothesis**  $H_0$ : **cancer = healthy** and select **one-sided hypothesis**  $H_1$ : **cancer < healthy**

# Two types of errors

	decide $H_0$	decide $H_1$
true $H_0$ probability	Correct action $1 - \alpha$	Type I error $\alpha$
true $H_1$ probability	Type II error $\beta$	Correct action power = $1 - \beta$

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

Sometimes the **type I error probability  $\alpha$**  is called the **significance level**, or the  **$\alpha$ -error**

**Instructions:** get  $\alpha$  from your boss or PI (e.g., 5% or 1%)

Prob( $H_0$  is true given the sample data)  $< \alpha$   
→ reject  $H_0$  and accept  $H_1$

Prob( $H_0$  is true given the sample data)  $> \alpha$   
→ accept  $H_0$  and reject  $H_1$

Type II error is much harder to estimate. Will deal with it later

# P-Values of Hypothesis Tests

- **P-value**: what is the probability to get the observed value of sample mean of  $\bar{x} = 48$  mRNAs/cell (or even smaller) and  $\sigma = 10$  mRNAs/cell in a healthy population with  $\mu_h = 50$  mRNAs/cell
- If **P-value is small** – the null hypothesis is likely wrong and thus, the **probability of making a type I error** (incorrectly rejecting the null hypothesis) **is small**
- P-value answers the question: if I reject the null hypothesis  $H_0$  based on the sample, what is the probability that I am making a type I error?

# P-Value vs $\alpha$ in Hypothesis Testing

- Problem with using a predefined  $\alpha$ : you **don't know by how much you exceeded it**
- Another approach is to calculate **Prob( $H_0$  is true given the sample data)** referred to as **P-value**.  
It is the smallest  $\alpha$  that would lead to rejection of null hypothesis
- You give your boss the P-value and let him/her decide if it is good enough
- Routinely with big datasets in genomics and systems biology P-values can be  $10^{-\text{large number}} \sim 10^{-100}$ . This number is used to judge the quality of the hypothesis



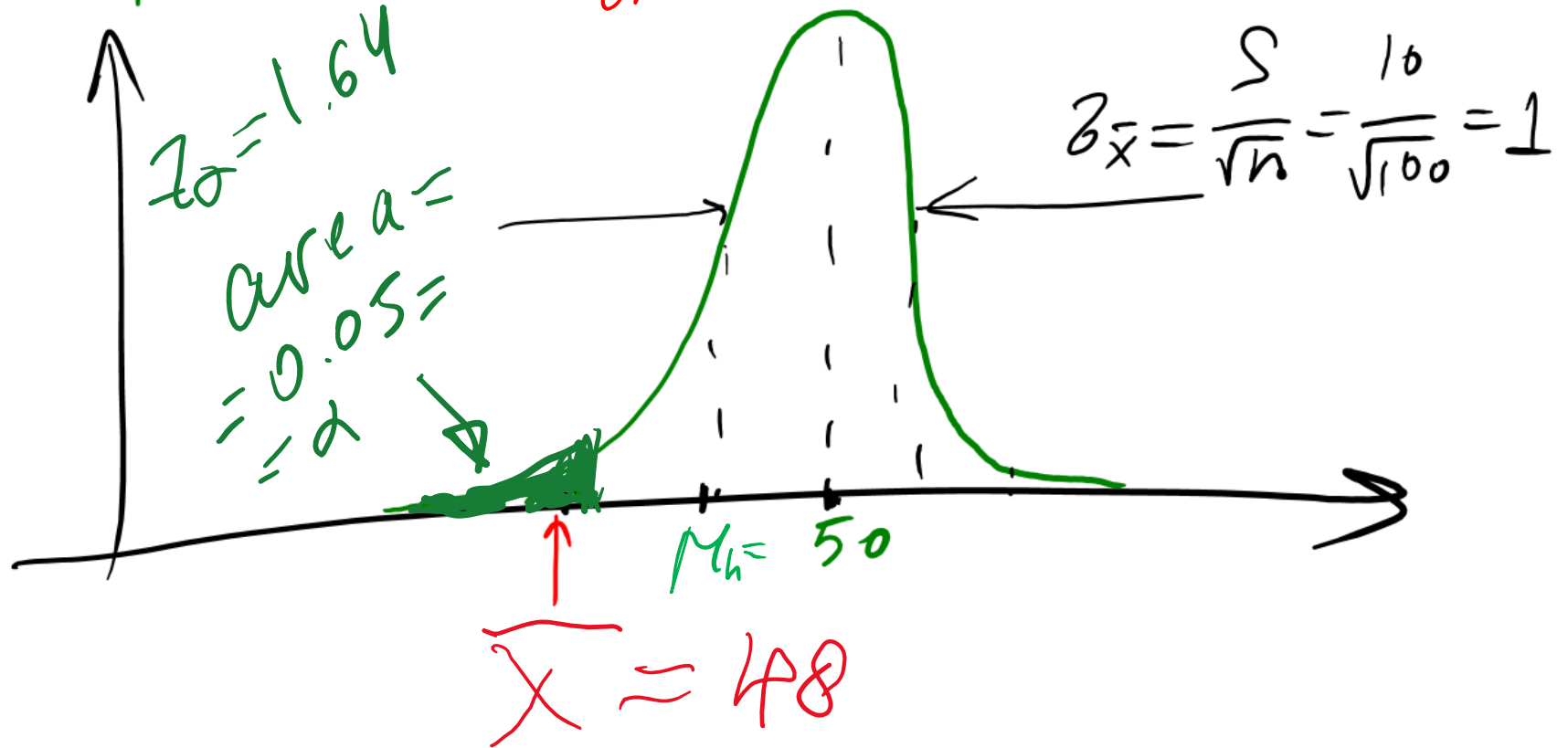


$$\mu_R = 50$$

$$H_0: \mu_C = \mu_R$$

$$n = 100, \bar{X} = 48, S = 10$$

One-sided hypothesis  $H_1: \mu_C < \mu_R$



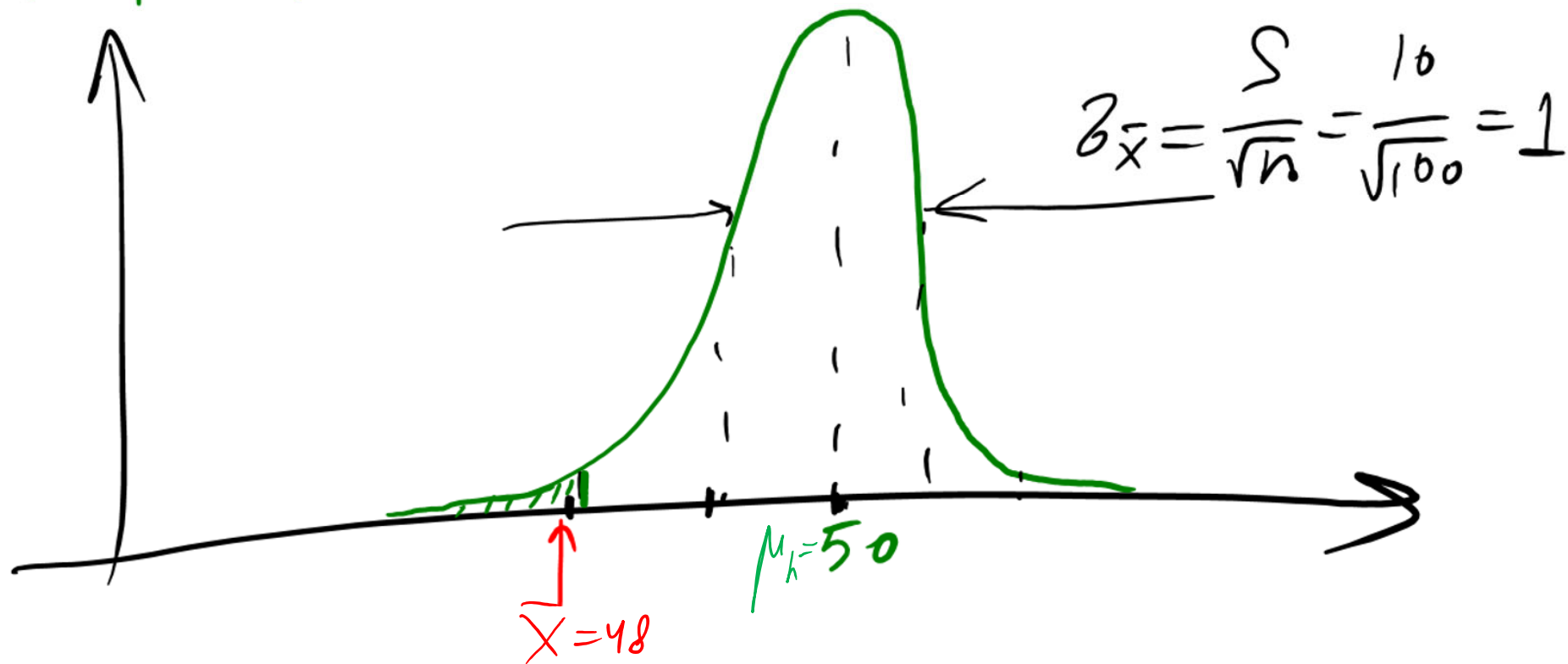
$$\text{P-value} = \text{Prob}(\bar{X}_n < 48 | H_0) =$$
$$\approx 2.5\%$$

$$\mu_h = 50$$

$$H_0: \mu_c = \mu_h$$

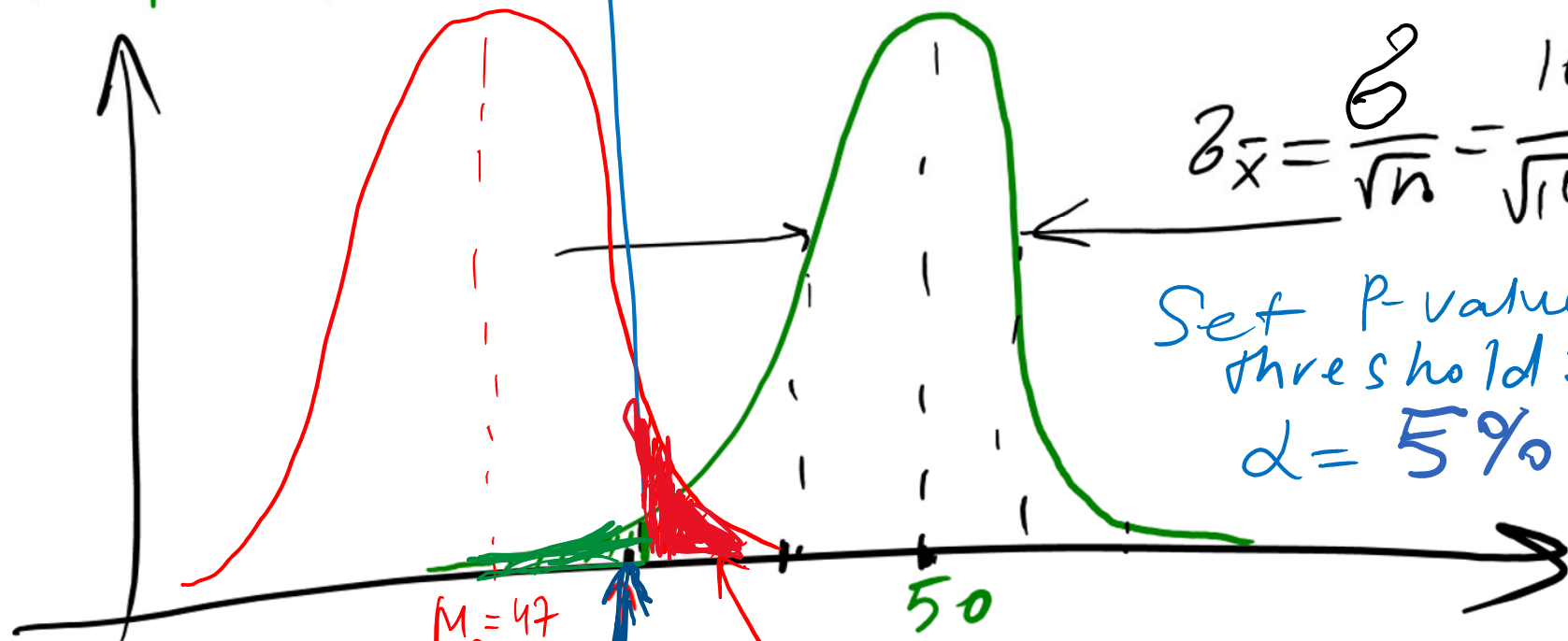
$$n = 100, \bar{X} = 48, S = 10$$

$$H_1: \mu_c < \mu_h$$



$\mu_h = 50$   
 $H_0: \mu_c = \mu_h$

$n = 100, \bar{X} = 48, \sigma = 10$   
 $H_1: \mu_c < \mu_h$



$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

Set P-value threshold:  
 $\alpha = 5\%$

$$\mu_h - z_{\alpha} \sigma_{\bar{x}} = 50 - 1.64 = 48.36$$

$$\beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) =$$

$$\alpha = 1 - \Phi(1.64) = 5\%$$

Type II error

$$\int_{48.36}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-47)^2}{2}\right) dx =$$

$$= 1 - \Phi(1.36) = 8.8\%$$

# Generalizations

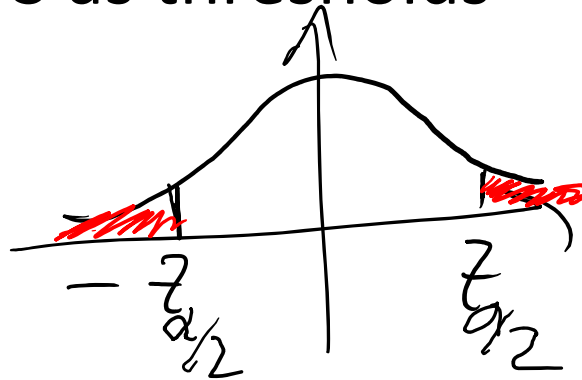
- What if  $H_1$  is a two-sided hypothesis?

- A: P-value is  $2(1-\Phi(|Z|))$ , where  $Z=(\bar{X}-\mu_0)/[S/\sqrt{n}]$

Compare it to: For one sided  $\mu_1 > \mu_0$  it is  $1-\Phi(Z)$

For one sided  $\mu_1 < \mu_0$  it is  $\Phi(Z)$

- If  $\alpha$  is given, use  $\mu_0 \pm z_{\alpha/2} * S$  as thresholds to reject the null hypothesis



- What if the sample size  $n$  is small (say  $n < 10$ ):

- A: Use t-distribution with  $n-1$  degrees of freedom for 2-sided  $P\text{-value} = 2(1 - \text{CDF\_Tdist}(|T|))$

where  $T = (\bar{X} - \mu_0) / [S / \sqrt{n}]$ .

- For a given  $\alpha$  use  $\mu_0 \pm t_{\alpha/2, n-1} T$  to reject the null hypothesis

# Type II Error and Choice of Sample Size

Assume you know the minimum  $\delta = |\mu_1 - \mu_0|$  that you care about.

What is the minimal sample you should use to separate  $H_0$  and  $H_1$  hypotheses if your tolerance to type I and type II errors is  $\alpha$  and  $\beta$  ?

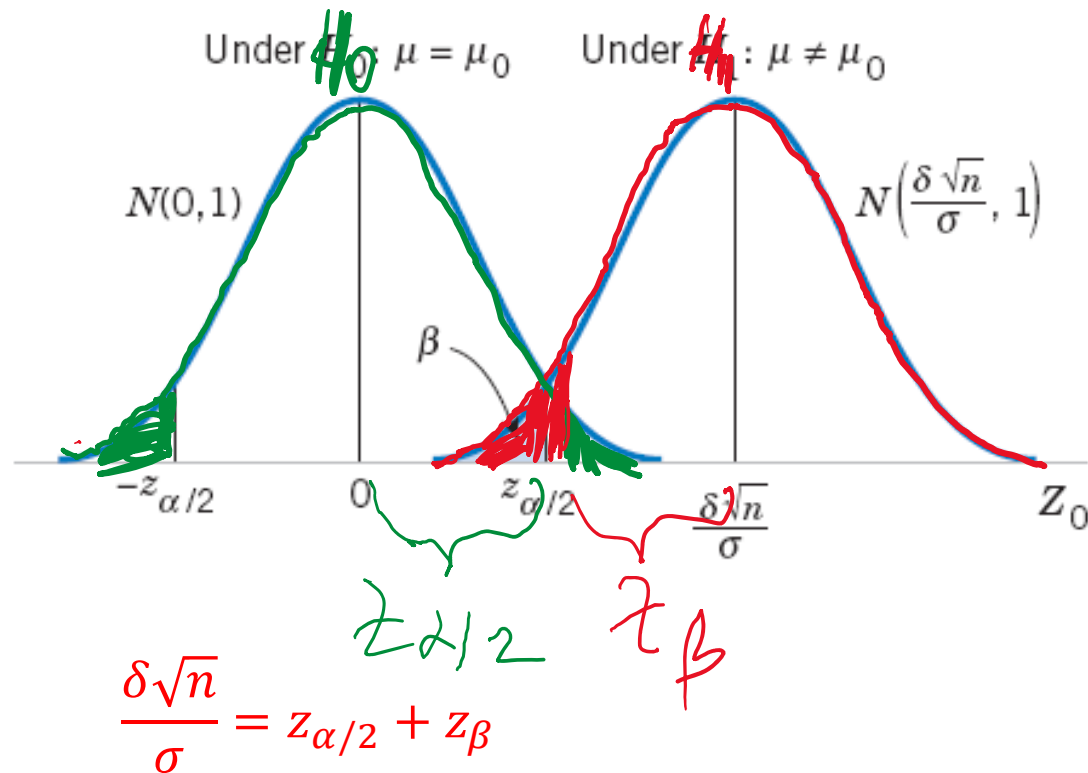


Figure 9-9 The distribution of  $Z_0$  under  $H_0$  and  $H_1$ .

$$n \approx \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0 \quad (9-22)$$

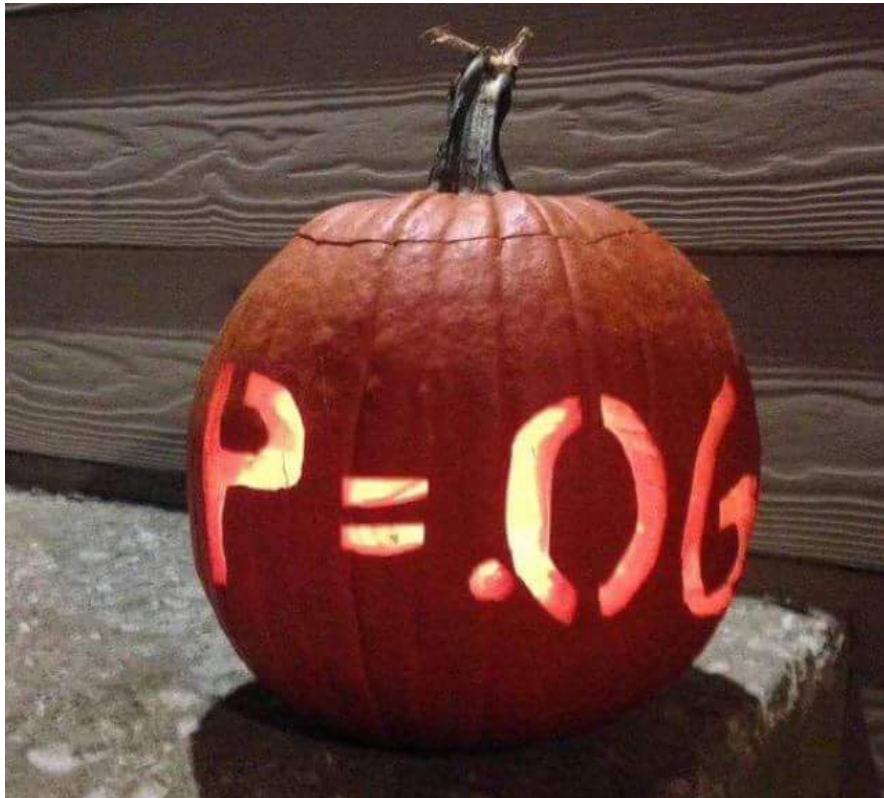
# Standard notation to indicate P-value with

**\*** , **\*\*** , **\*\*\***

Table 11.1: A commonly adopted convention for reporting  $p$  values: in many places it is conventional to report one of four different things (e.g.,  $p < .05$ ) as shown below. I've included the "significance stars" notation (i.e., a \* indicates  $p < .05$ ) because you sometimes see this notation produced by statistical software. It's also worth noting that some people will write *n.s.* (not significant) rather than  $p > .05$ .

Usual notation	Signif. stars	English translation	The null is...
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$ .	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$ .	Rejected
$p < .001$	***	The test was significant at all levels	Rejected

---



Happy  
Halloween!  
(belated)

Credit: Trust me,  
I'm a "Biologist"  
Facebook community

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	] — HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	] — SIGNIFICANT
0.049	
0.050	] — OH CRAP. REDO CALCULATIONS.
0.051	] — ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	] — HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	] — HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
$\geq 0.1$	

Credit: XKCD  
comics



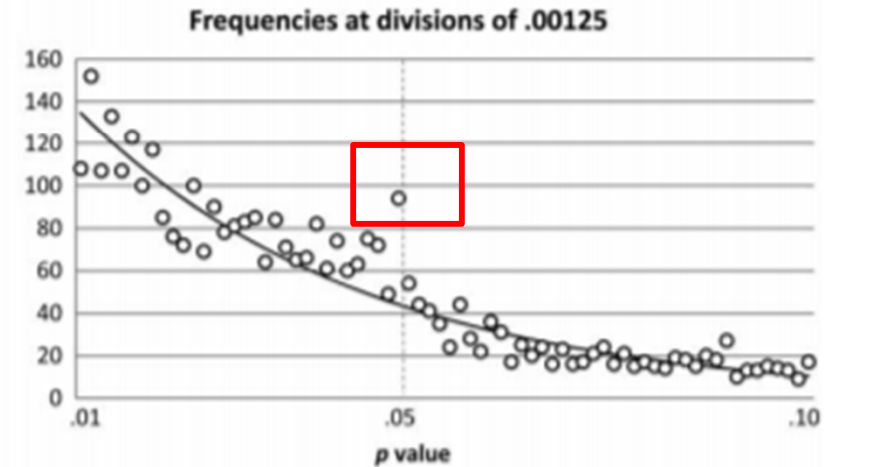
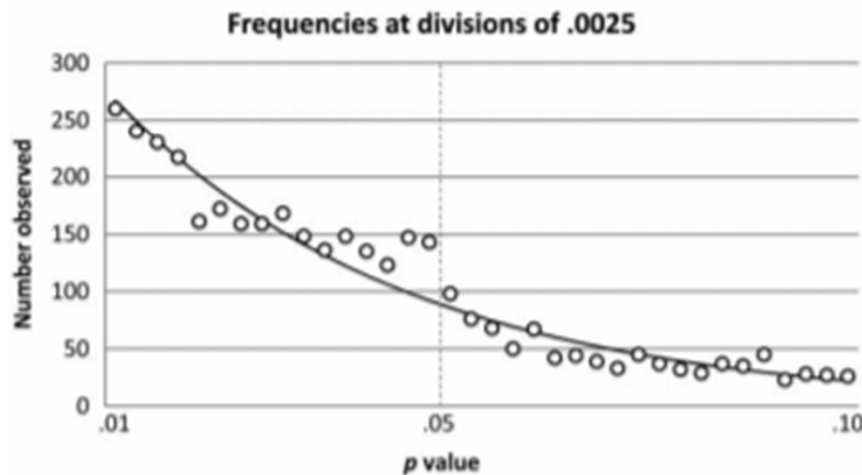
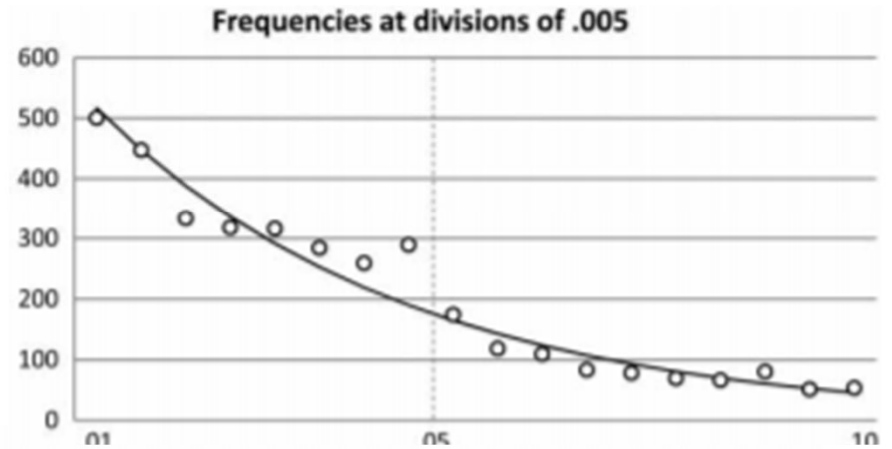
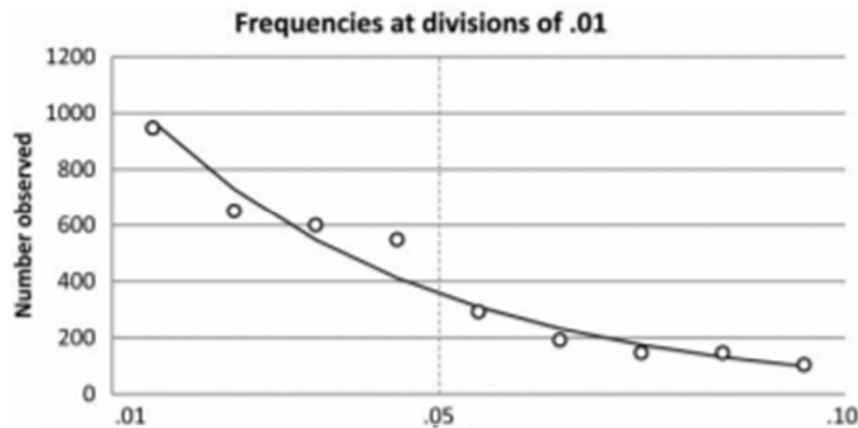
# A peculiar prevalence of $p$ values just below .05

E. J. Masicampo<sup>1</sup>, and Daniel R. Lalande<sup>2</sup>

<sup>1</sup>Department of Psychology, Wake Forest University, Winston-Salem, NC, USA

<sup>2</sup>Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada

## MASICAMPO AND LALANDE





Credit: XKCD  
comics

# WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS  
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED  
WHY IS THERE A LINE THROUGH HTTPS  
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK  
WHY IS HTTPS IMPORTANT

# QUESTIONS FOUND IN GOOGLE AUTOCOMPLETE



WHY ARE THERE WEEKS  
WHY DO I FEEL DIZZY

WHY AREN'T ECONOMISTS RICH

WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN  
WHY IS THERE PHLEGM

WHY DO AMERICANS CALL IT SOCCER

WHY IS PSYCHIC WEAK TO BUG

WHY ARE MY EARS RINGING

WHY DO CHILDREN GET CANCER

WHY ARE THERE SO MANY AVENGERS

WHY IS POSEIDON ANGRY WITH ODYSSEUS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS THERE ICE IN SPACE

# WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED  
WHY IS SPACE BLACK  
WHY IS OUTER SPACE SO COLD  
WHY ARE THERE PYRAMIDS ON THE MOON  
WHY IS NASA SHUTTING DOWN



WHY IS THERE AN OWL IN MY BACKYARD  
WHY IS THERE AN OWL OUTSIDE MY WINDOW  
WHY IS THERE AN OWL ON THE DOLLAR BILL  
WHY DO OWLS ATTACK PEOPLE  
WHY ARE AK 47s SO EXPENSIVE  
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE  
WHY ARE THERE GODS  
WHY ARE THERE TWO SPOCKS

WHY ARE DOGS AFRAID OF FIREWORKS  
WHY IS THERE NO KING IN ENGLAND

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY DO SPIDERS COME INSIDE

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY DO SPIDER BITES ITCH

WHY IS DYING SO SCARY

WHY IS THERE NO GPS IN LAPTOPS

WHY DO KNEES CLICK

WHY AREN'T THERE E GRADES

WHY IS SEX SO IMPORTANT



WHY IS MT VESUVIUS THERE

WHY DO THEY SAY T MINUS

WHY ARE THERE OBELISKS

WHY ARE WRESTLERS ALWAYS WET

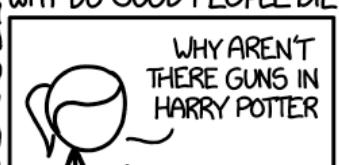
WHY ARE OCEANS BECOMING MORE ACIDIC

WHY IS ARWEN DYING

WHY AREN'T MY QUAIL LAYING EGGS

WHY AREN'T MY QUAIL EGGS HATCHING

WHY ARE CIGARETTES LEGAL  
WHY ARE THERE DUCKS IN MY POOL  
WHY IS JESUS WHITE  
WHY IS THERE LIQUID IN MY EAR  
WHY DO Q TIPS FEEL GOOD  
WHY DO GOOD PEOPLE DIE



WHY ARE ULTRASOUNDS IMPORTANT  
WHY ARE ULTRASOUND MACHINES EXPENSIVE  
WHY IS STEALING WRONG

WHY IS LIFE SO BORING

WHY DO WHALES JUMP  
WHY ARE WITCHES GREEN  
WHY ARE THERE MIRRORS ABOVE BEDS

WHY DO I SAY UH  
WHY IS SEA SALT BETTER  
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS

WHY IS THERE NOT A POKEMON MMO  
WHY IS THERE LAUGHING IN TV SHOWS  
WHY ARE THERE DOORS ON THE FREEWAY

WHY ARE THERE SO MANY SVCHOST.EXE RUNNING  
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA  
WHY ARE THERE SCARY SOUNDS IN MINECRAFT

WHY IS THERE KICKING IN MY STOMACH  
WHY ARE THERE TWO SLASHES AFTER HTTP  
WHY ARE THERE CELEBRITIES

WHY DO SNAKES EXIST  
WHY DO OYSTERS HAVE PEARLS  
WHY ARE DUCKS CALLED DUCKS

WHY DO THEY CALL IT THE CLAP  
WHY ARE KYLE AND CARTMAN FRIENDS  
WHY IS THERE AN ARROW ON AANG'S HEAD

WHY ARE TEXT MESSAGES BLUE  
WHY ARE THERE MUSTACHES ON CLOTHES  
WHY ARE THERE MUSTACHES ON CARS

WHY ARE THERE MUSTACHES EVERYWHERE  
WHY ARE THERE SO MANY BIRDS IN OHIO  
WHY IS THERE SO MUCH RAIN IN OHIO

WHY IS OHIO WEATHER SO WEIRD  
WHY ARE THERE MALE AND FEMALE BIKES  
WHY ARE THERE BRIDESMAIDS

WHY DO DYING PEOPLE REACH UP  
WHY AREN'T THERE VARICOSE ARTERIES  
WHY ARE OLD KUNGONS DIFFERENT

WHY ARE THERE SQUIRRELS  
WHY IS PROGRAMMING SO HARD  
WHY IS THERE A 0 OHM RESISTOR  
WHY DO AMERICANS HATE SOCCER  
WHY DO RHYMES SOUND GOOD

WHY DO TREES DIE  
WHY IS THERE NO SOUND ON CNN  
WHY AREN'T POKEMON REAL  
WHY AREN'T BULLETS SHARP  
WHY DO DREAMS SEEM SO REAL

WHY AREN'T THERE DINOSAUR GHOSTS

WHY ARE THERE FEMALE MR NIMES

WHY IS GPS FREE



WHY IS THERE HELL IF GOD FORGIVES

WHY ARE THERE SQUIRRELS