Confidence intervals for the population variance σ^2 based on the sample variance s^2

Confidence interval for the population variance σ^2

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

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

- How to determine the confidence interval?

$\vec{X} = (x_1, x_2, \ldots, x_n)$ $\quad x_i \to x_i - \bar{x}$

$y = |\vec{x}|^2 = \sum x_i^2 = (n-1)s^2$ $\quad \sum_{i=1}^{n} x_i = 0$

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

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

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

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

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

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

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

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

Definition (Eq. 8-17)

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

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

has a chi-square (χ^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 σ^2

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

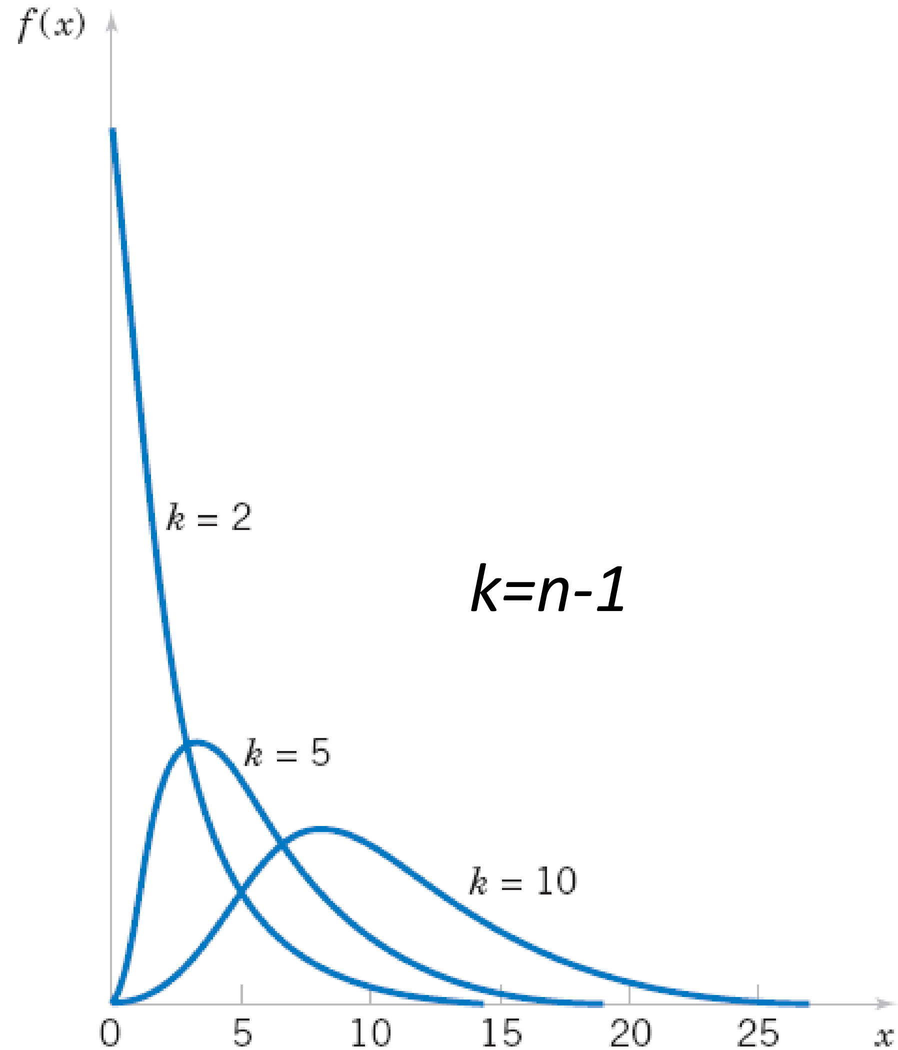
It is just Gamma PDF with $r=(n-1)/2$, and $\lambda=1/2$

Mean value:

$n-1$

Standard deviation:

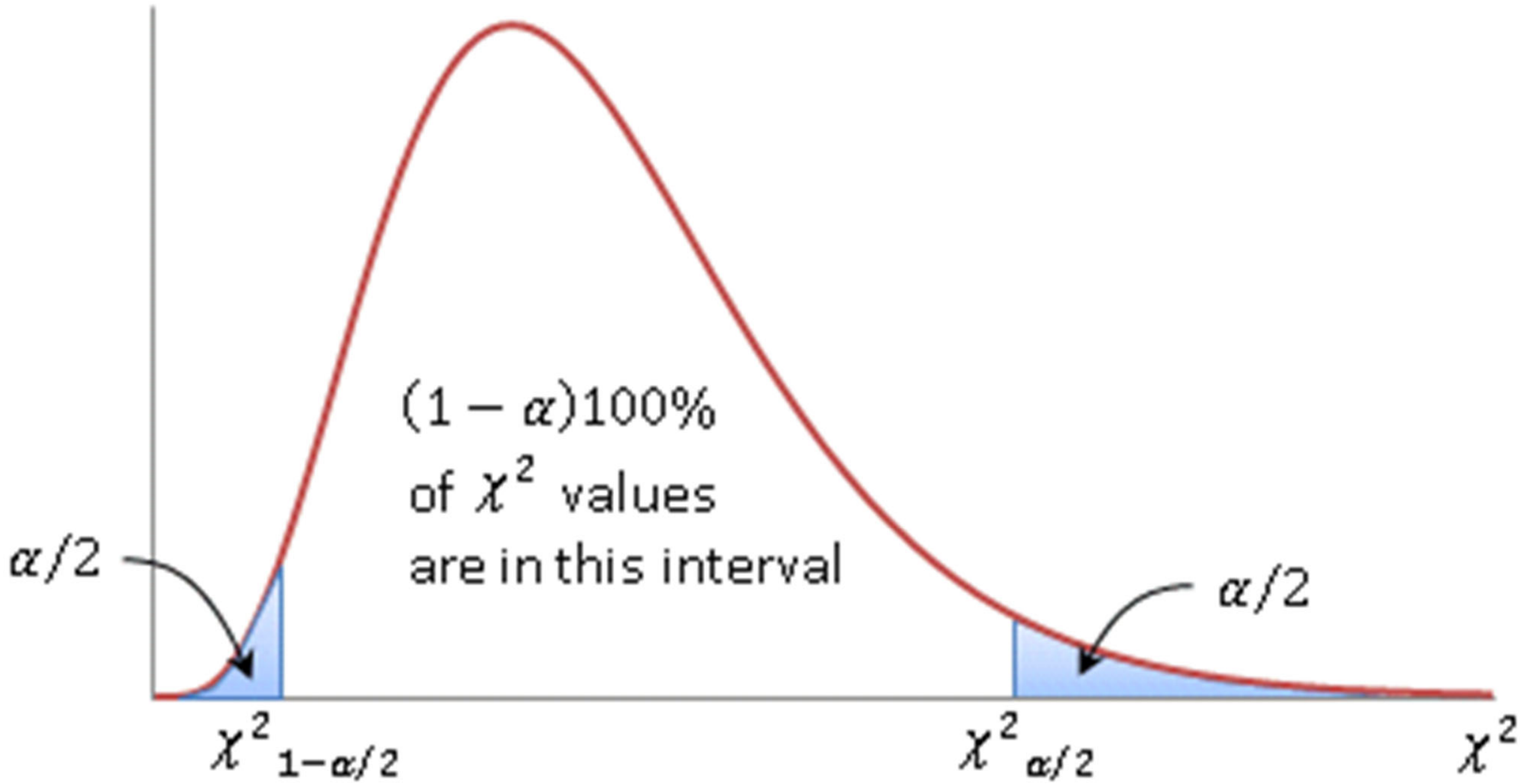
$$\sqrt{2(n-1)}$$



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

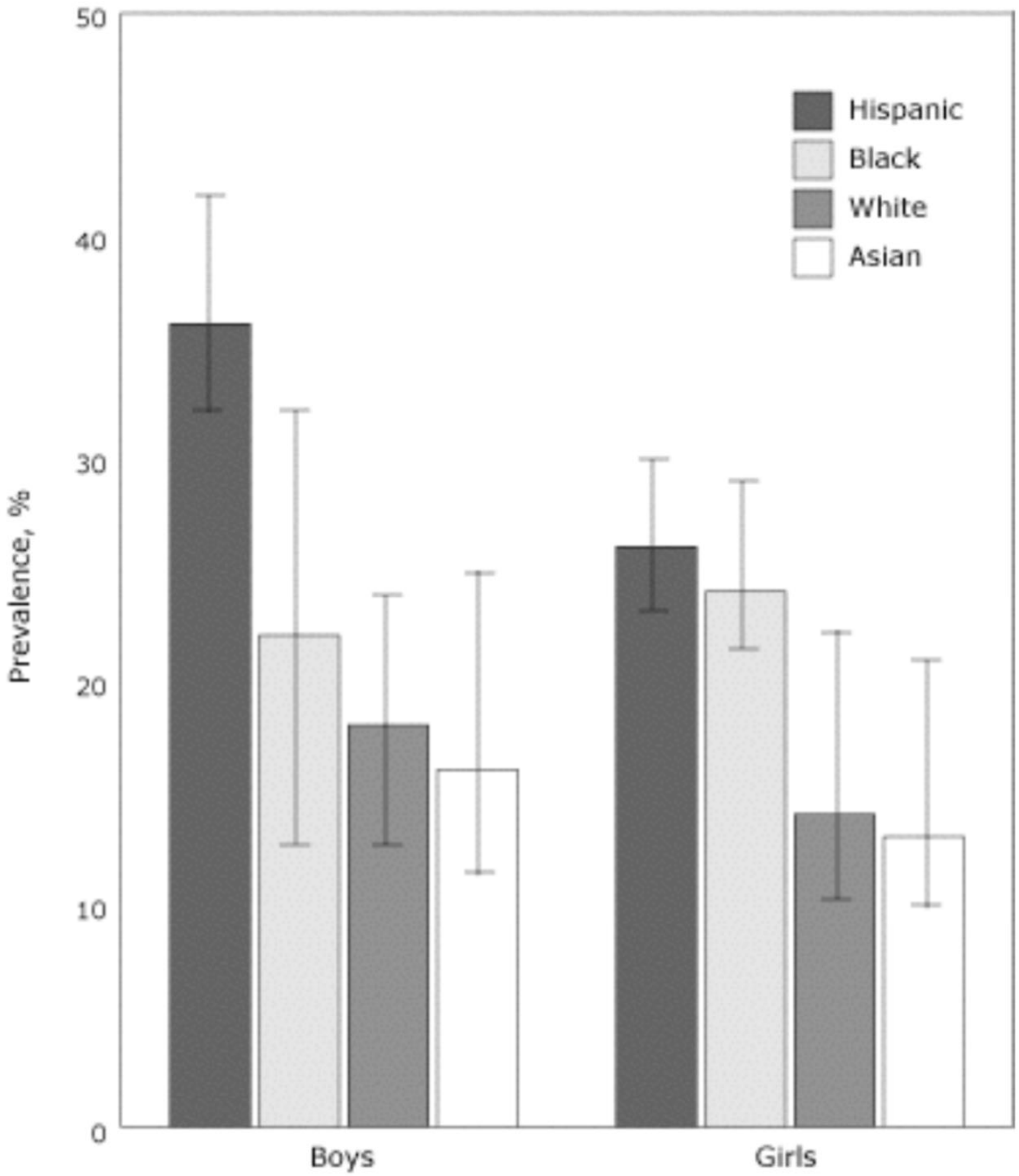
(Eq. 8-19)

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then **a 100(1 − α)% confidence interval on σ^2 is**

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{8-19}$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ 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 σ** 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 *$\hat{p}$=(# of obese kids in the sample)/(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

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

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

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

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® 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® Chocolate Candies is:

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

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

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

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

M&M'S® 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 × 7oz bags with 210 candies each

For red M&Ms $p = 0.13$

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

~1 × 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 **<u>lowered</u>** in population of cancer patients compared to the healthy population.

- We know that mean gene expression in the healthy population is μ_h=50 mRNAs/cell We are interested in deciding whether or not the mean expression in cancer population is **<u>lower than</u>** in healthy population. Let's call hypothesis H_1. Here H_1 is **<u>one-sided</u>**

- If we asked: cancer is <u>not equal</u> to healthy H_1 would be a **<u>two-sided</u> hypothesis**

- Assume we have a sample of 100 cancer patients with sample mean $\bar{x}$ =48 mRNAs/cell and standard deviation σ=10 mRNA/cell

- Can we use our sample to reject the "business as usual" or **<u>null hypothesis</u> H_0**: cancer = healthy and select **<u>one-sided hypothesis</u> 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 α
true H_1 probability	Type II error β	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 α** is called the **significance level,** or the **α-error**

Instructions: get α from your boss or PI (e.g., 5% or 1%)

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

Prob(H_0 is true given the sample data) > α
→ 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 σ=10 mRNAs/cell in a healthy population with μ_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 α in Hypothesis Testing

- Problem with using a predefined α: 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 the smallest α 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 $10^{\text{-large number\textasciitilde10-100}}$. This number is used to judge the quality of the hypothesis

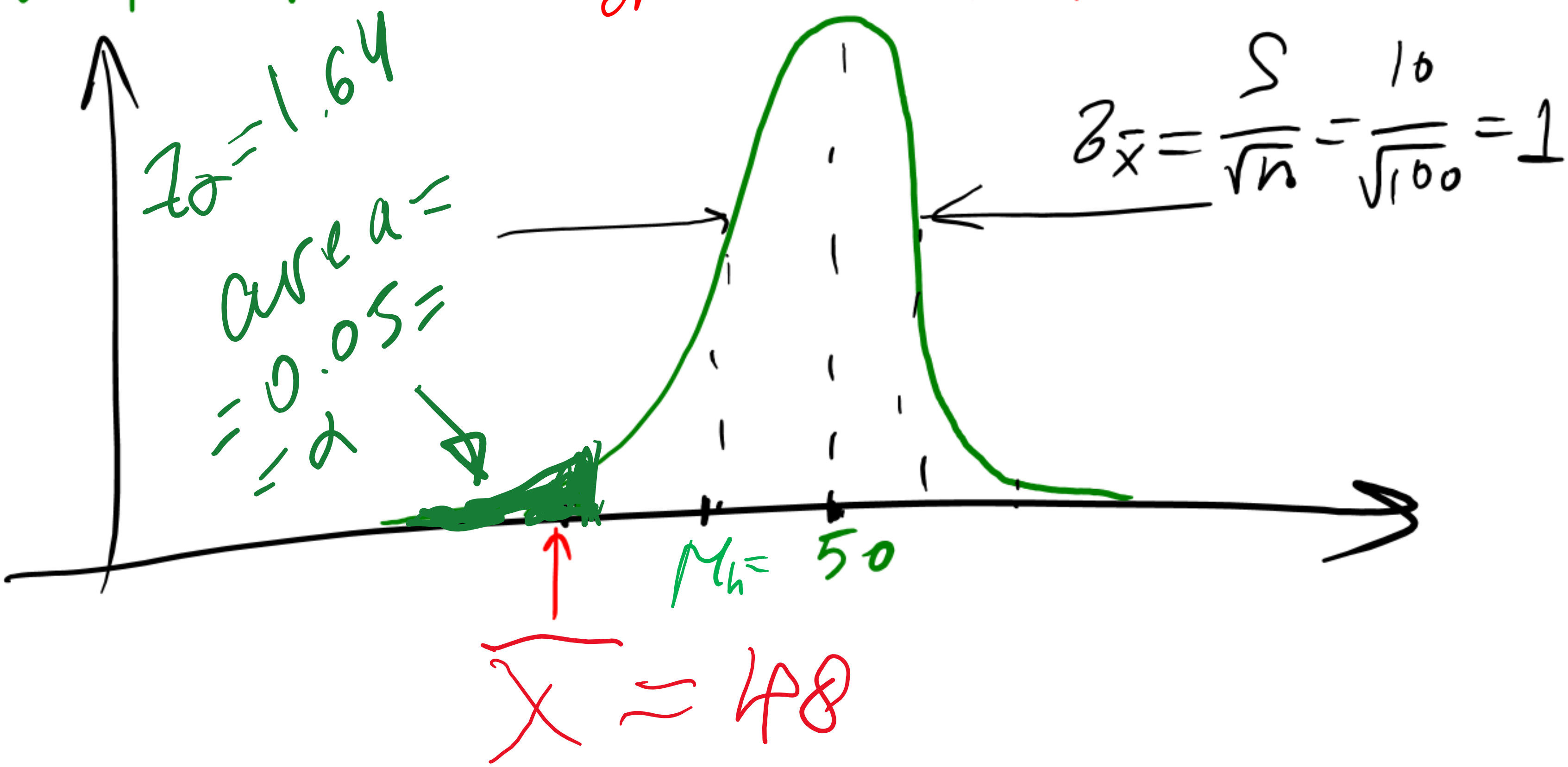
$\mu_h = 50$

$H_0: \mu_c = \mu_h$

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

One-sided hypothesis

$H_1: \mu_c < \mu_h$



$$\text{P-value} = \text{Prob}(\bar{X}_h < 48 \mid H_0) = \cong 2.5\%$$

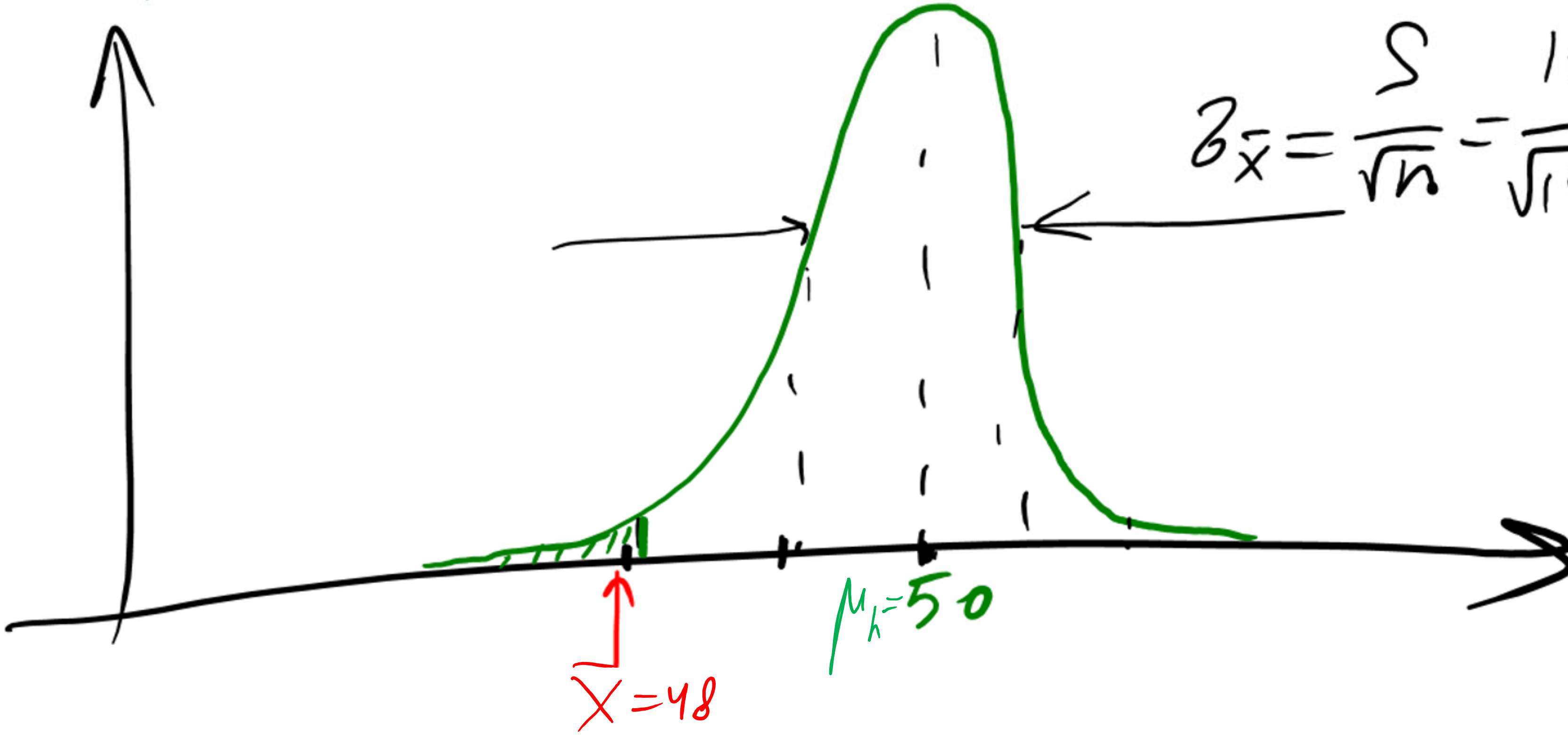
$\mu_h = 50$

$H_0: \mu_c = \mu_h$

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

$H_1: \mu_c < \mu_h$

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



$\mu_h = 50$

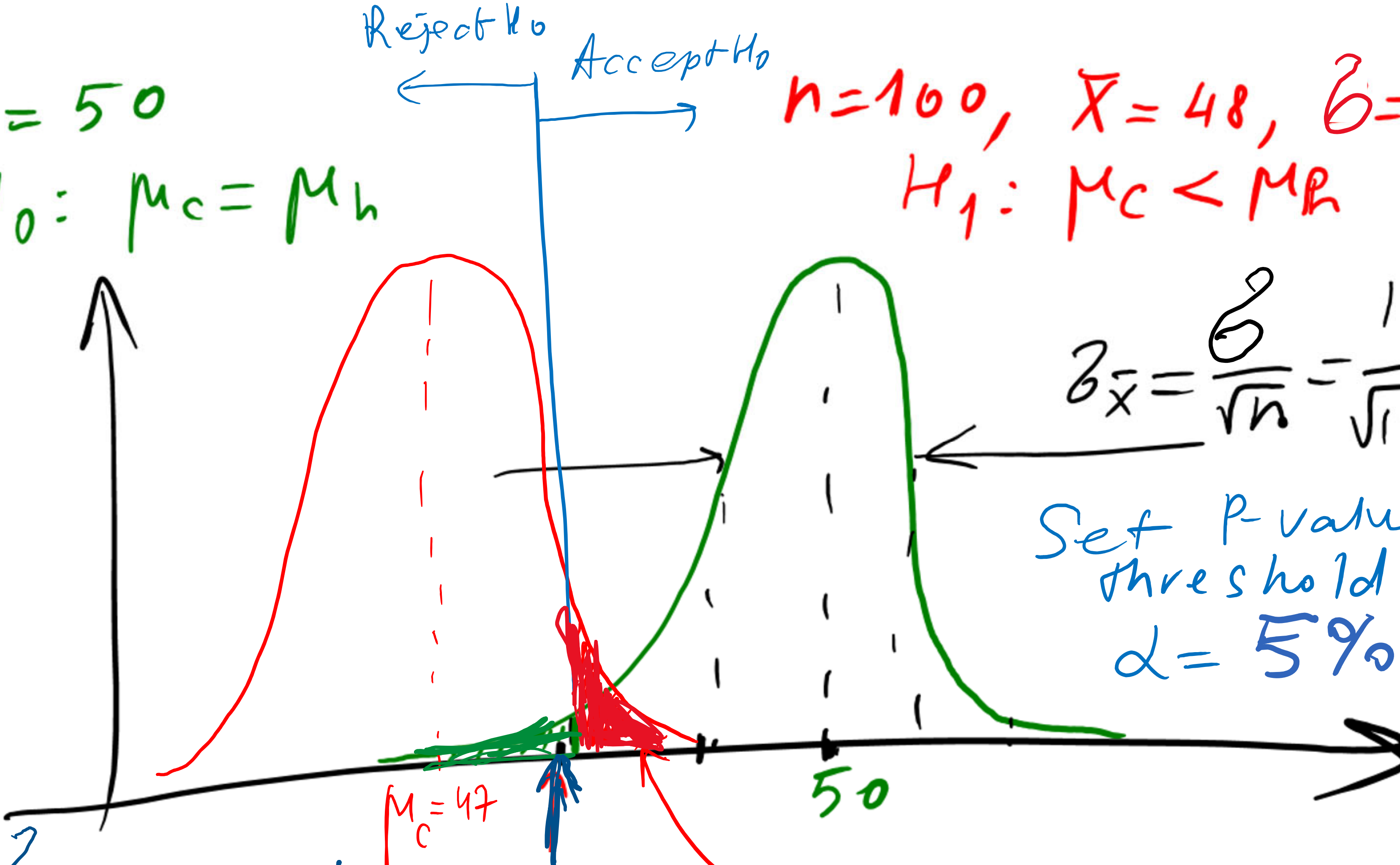
$H_0: \mu_c = \mu_h$

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

$H_1: \mu_c < \mu_h$

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

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



$\mu_h - z_\alpha \frac{\sigma}{\sqrt{n}} = 50 - 1 \cdot 1.64 = 48.36$

$\beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = \int_{48.36}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-47)^2}{2}\right) dx = 1 - \Phi(1.36) = 8.8\%$

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

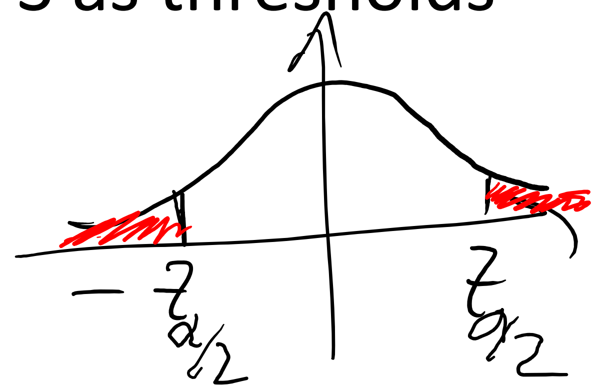
Generalizations

- 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 sized $\mu_1 > \mu_0$ it is $1-\Phi(Z)$

  For one sized $\mu_1 < \mu_0$ it is $\Phi(Z)$
- If α is given, use μ_0 +/-$z_{\alpha/2}$*S as thresholds

to reject the null hypothesis

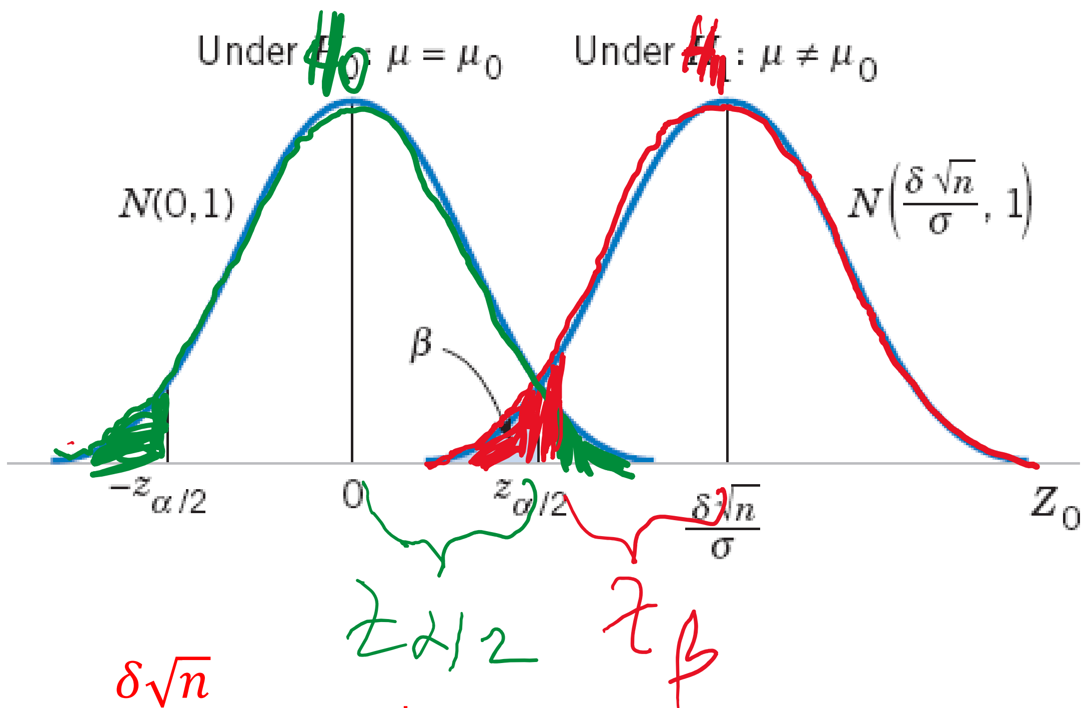


- What if the sample size n is small (say n<10):
- A: Use t-distribution with n-1 degrees of freedom for 2-sided $P\text{-}value=2(1-CDF\_Tdist(|T|))$ where $T=(\bar{X}-\mu_0)/[S/\sqrt{n}]$.
- For a given α use $\mu_0 +/- t_{\alpha/2,n-1} T$ to reject the null hypothesis

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 H0 and H1 hypotheses if your tolerance to type I and type II errors is α and β ?

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



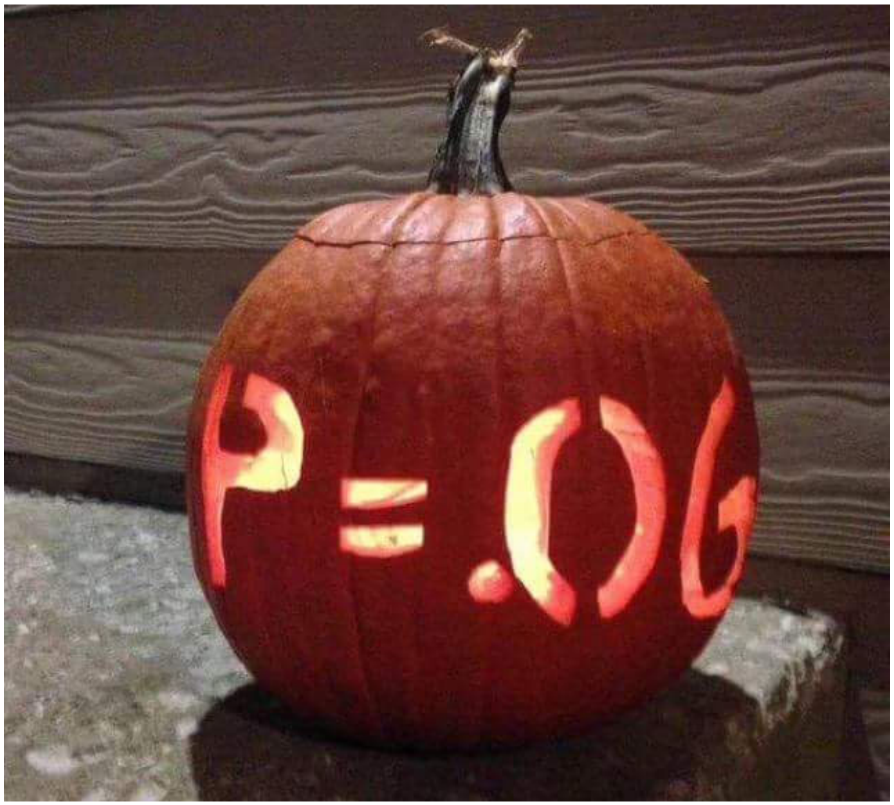
$$\frac{\delta\sqrt{n}}{\sigma} = z_{\alpha/2} + z_\beta$$

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

Standard notation to indicate P-value with *, **, ***

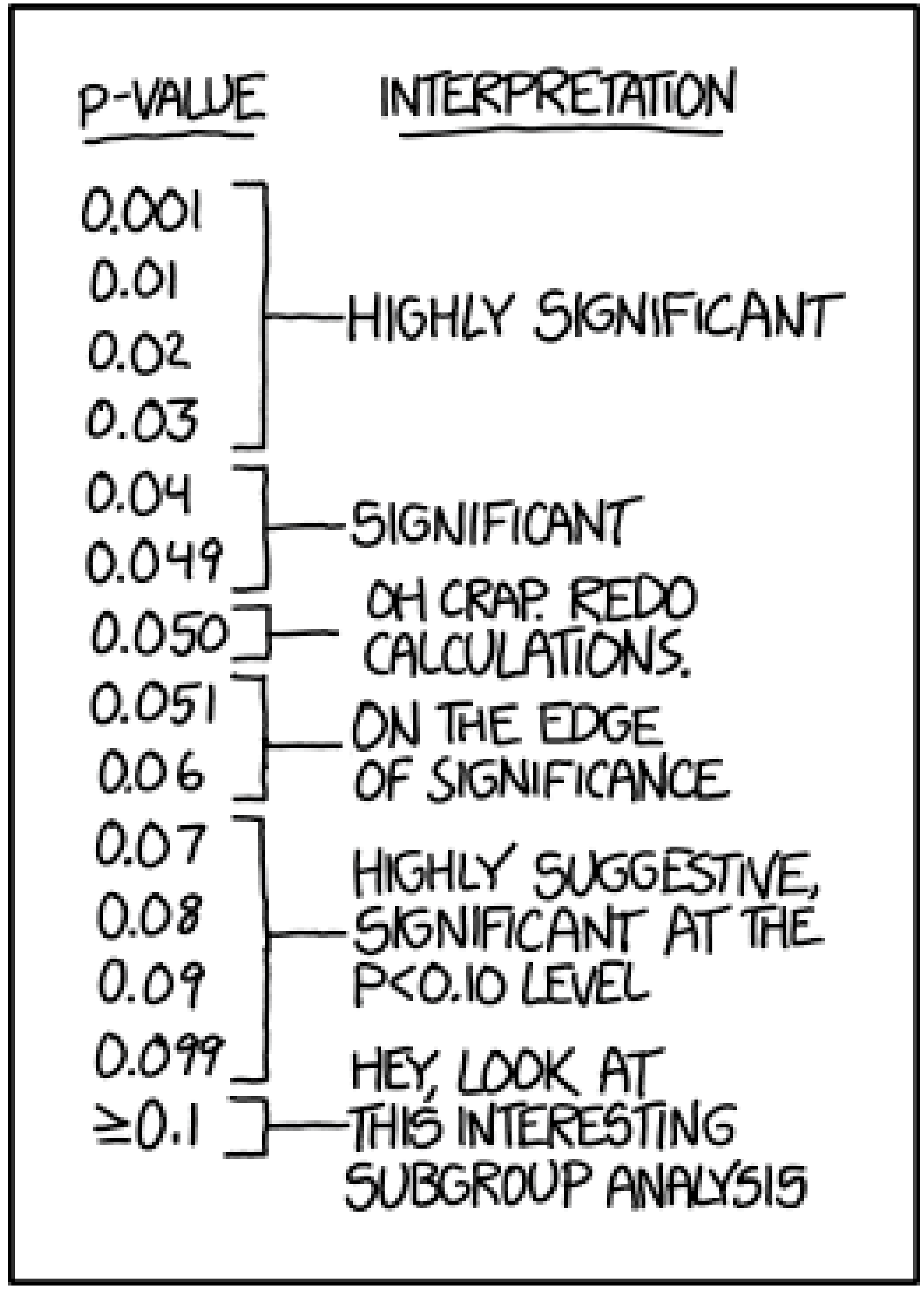
Table 11.1: A commonly adopted convention for reporting p values: in many places it is conventional to report one of four different things (e.g., $p < .05$) as shown below. I've included the "significance stars" notation (i.e., a * indicates $p < .05$) because you sometimes see this notation produced 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



Credit: XKCD
comics

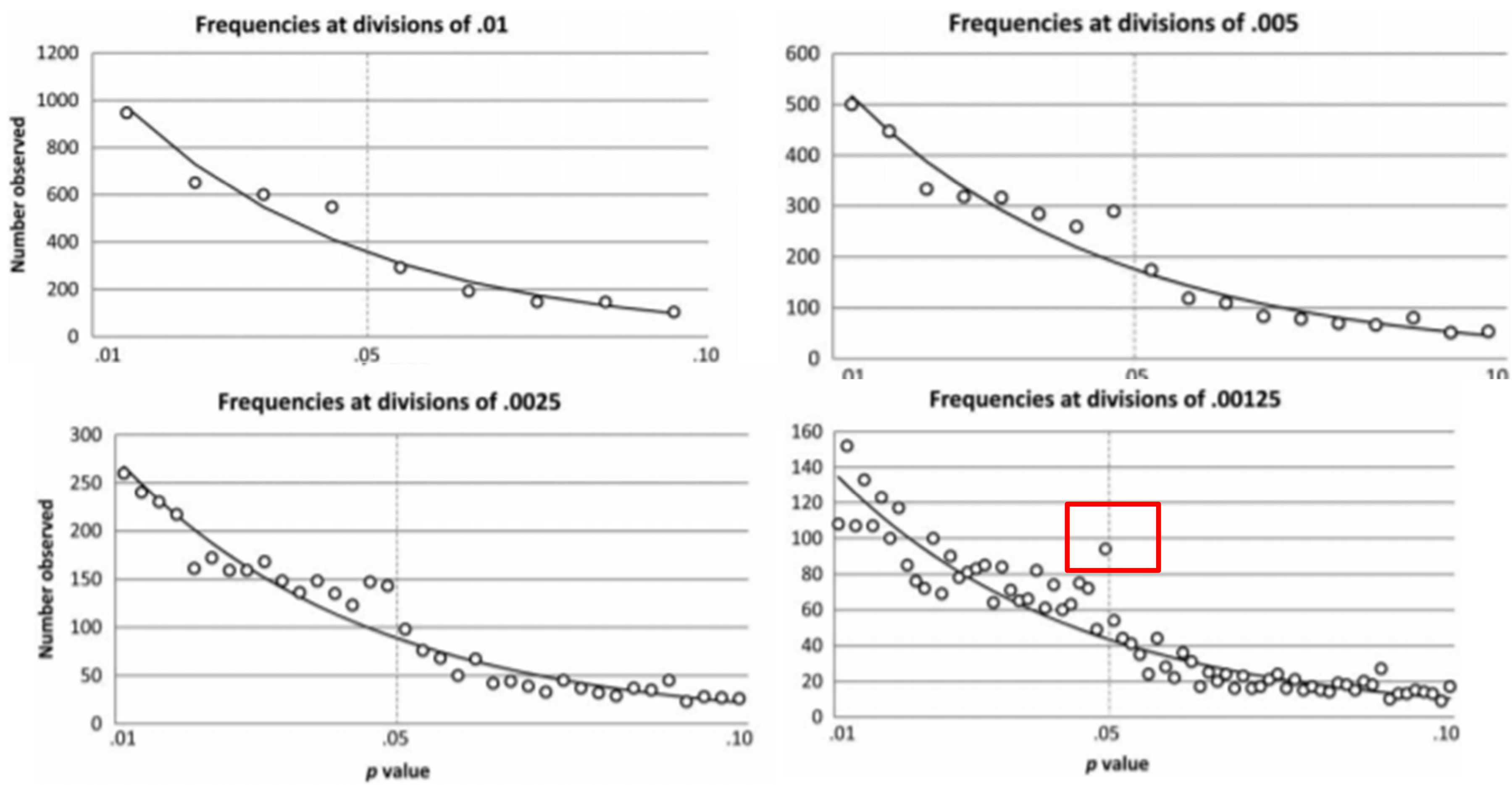
THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2012, 65 (11), 2271–2279

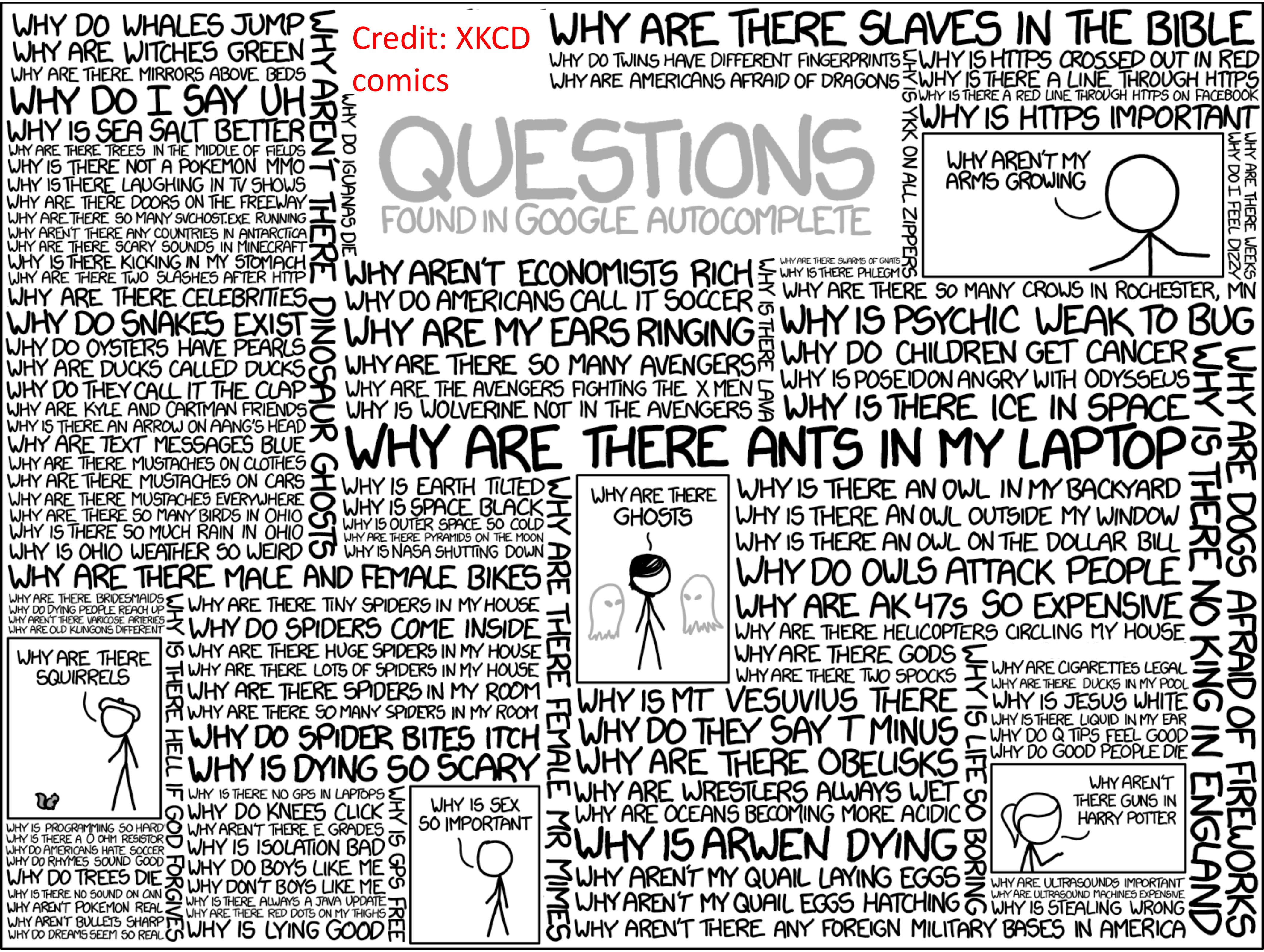
Psychology Press
Taylor & Francis Group

A peculiar prevalence of *p* values just below .05

E. J. Masicampo[1], and Daniel R. Lalande[2]

[1]Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
[2]Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada





Credit: XKCD comics