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 $\mu_h=50$ 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 H₁. Here H₁ is <u>one-sided</u>
- If we asked: cancer is <u>not equal</u> to healthy H₁ 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 $\sigma=$ 10 mRNA/cell
- Can we use our sample to reject the "business as usual" or null hypothesis H₀: cancer = healthy and select one-sided hypothesis H₁: cancer < healthy

Two types of errors

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

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

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

 \rightarrow reject H₀ and accept H₁

Prob(H_0 is true given the sample data) > α

→ accept H₀ and reject H₁

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 H₀ 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^{-large number~10-100}. This number is used to judge the quality of the hypothesis

n=100, X=48, S=10 Hj: Mc < MR Ho: Mc=Mh x = 48P-Value = Prob (Xh < 48) Hot-2.5%

$$M_{R} = 50$$
 $H_{0}: M_{C} = M_{h}$
 $M_{1}: M_{C} < M_{h}$
 $M_{2} = \frac{S}{\sqrt{n}} - \frac{16}{\sqrt{100}} = 1$
 $M_{R} = 50$
 $M_{1}: M_{C} < M_{h}$
 $M_{2} = \frac{S}{\sqrt{n}} - \frac{16}{\sqrt{100}} = 1$

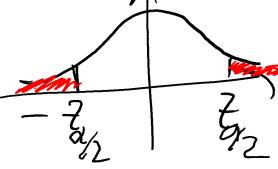
Keject 10 AcceptHo n=100, X=48, 6=10 H1: MC < MR MR = 50 Ho: Mc = Mh Set P-value threshold: d= 5% $\beta = P\left(A c \operatorname{cept} H_0 | H_1 \text{ is true}\right) = \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X-47)!}{2}\right) dX = \right)$ d= 1-9 (1.6) = 5% = 1-P(1.36=8.8% 48.36

Generalizations

- What if H₁ is a two-sided hypothesis?
- A: P-value is $2(1-\Phi(|Z|))$, where $Z=(\overline{X}-\mu_0)/[S/\sqrt{n}]$ 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 $\mu_0 + /-z_{\alpha/2} *S$ 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-value=2(1- $CDF_Tdist(|T|))$ where T= $(\bar{X}$ - $\mu_0)/[S/\sqrt{n}]$.
- For a given α use μ_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 ?

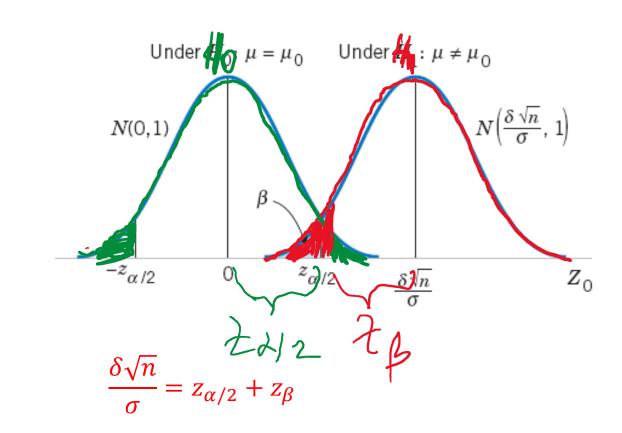


Figure 9-9 The distribution of Z_0 under H_0 and H_1 .

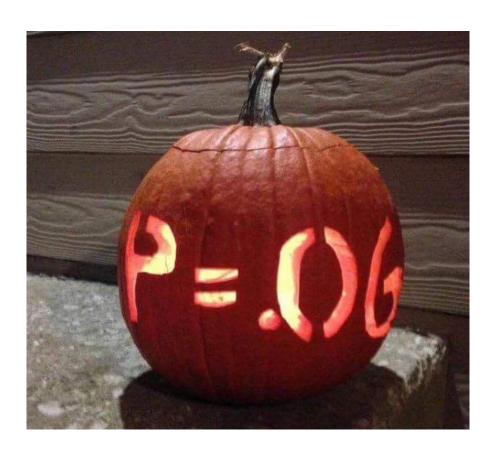
$$n \simeq \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2}$$
 where $\delta = \mu - \mu_0$ (9-22)

Standard notation to indicate P-value with

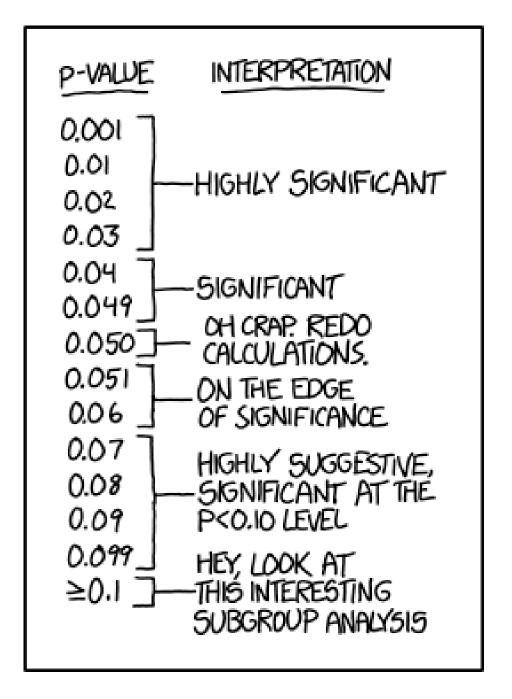
* ** ***

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



Credit: XKCD

comics

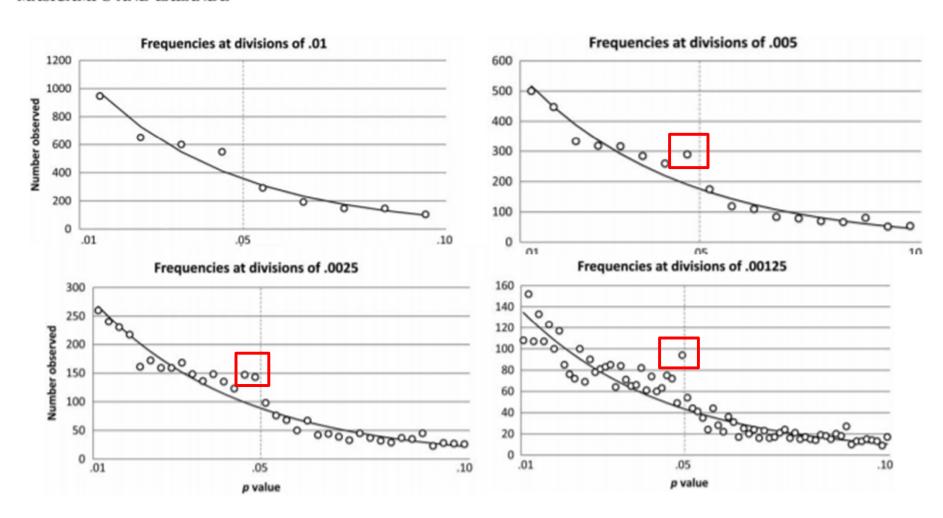


A peculiar prevalence of p values just below .05

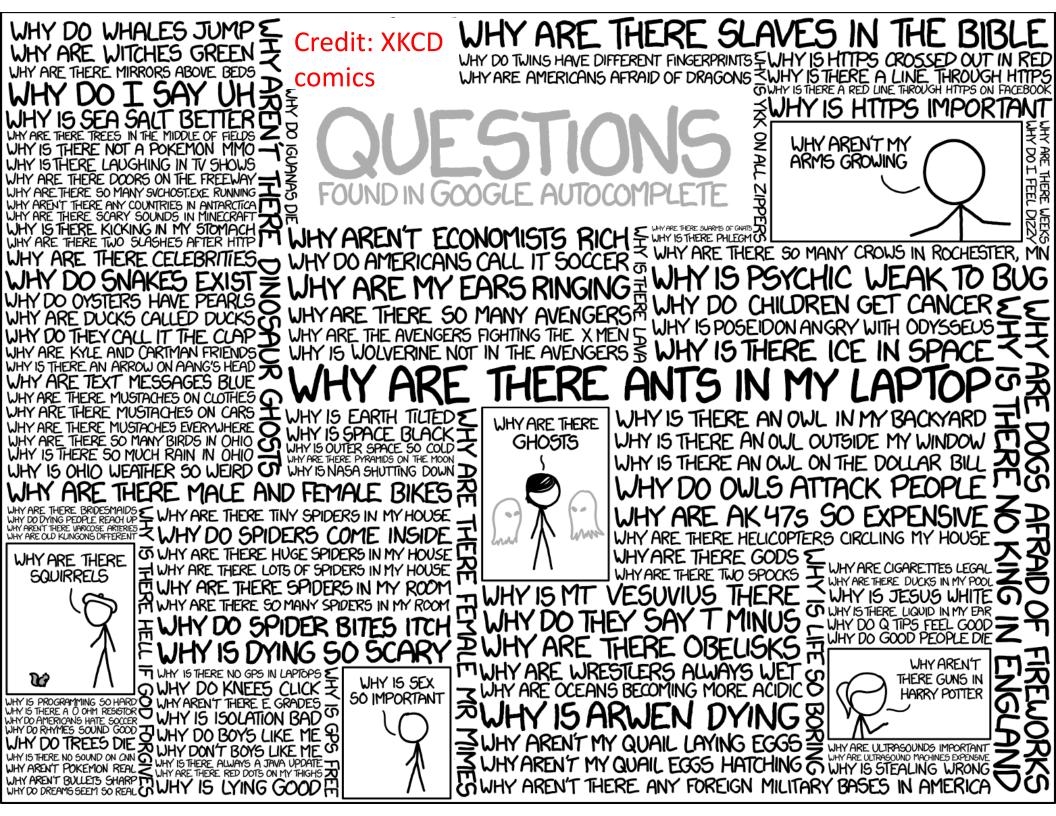
E. J. Masicampo¹, and Daniel R. Lalande²

¹Department of Psychology, Wake Forest University, Winston-Salem, NC, USA

MASICAMPO AND LALANDE



²Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada



Hypothesis testing: two samples

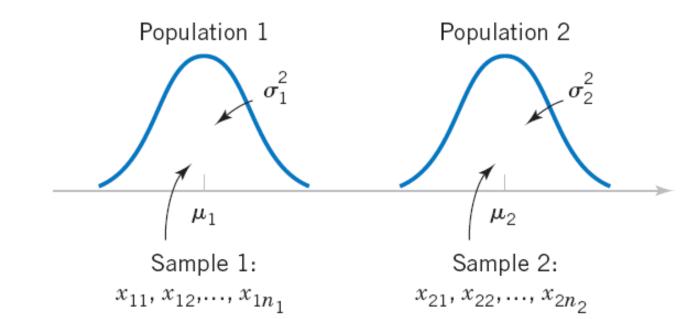


Figure 10-1 Two independent populations.

Figure 10-1 Two independent populations.

Assumptions

- 1. $X_{11}, X_{12}, \dots, X_{1n_1}$ is a random sample from population 1.
- 2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
- 3. The two populations represented by X_1 and X_2 are independent.
- **4.** Both populations are normal.

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

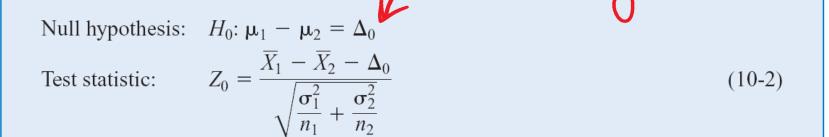
The quantity

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$
(10-1)

has a N(0, 1) distribution.

10-2.1 Hypothesis Tests for a Difference in Means,

Variances Known



Alternative Hypotheses	<i>P</i> -Value	for Fixed-Level Tests
H_1 : $\mu_1 - \mu_2 \neq \Delta_0$	Probability above $ z_0 $ and probability below $- z_0 $,	$z_0 > z_{\alpha/2} \text{ or } z_0 < -z_{\alpha/2}$
H_1 : $\mu_1 - \mu_2 > \Delta_0$	$P = 2[1 - \Phi(z_0)]$ Probability above z_0 ,	$z_0 > z_{\alpha}$
H_1 : $\mu_1 - \mu_2 < \Delta_0$	$P = 1 - \Phi(z_0)$ Probability below z_0 ,	$z_0 < -z_{\alpha}$
	$P = \Phi(z_0)$	

10-2.1 Hypotheses Tests on the Difference in Means, Variances Unknown

Case 2:
$$\sigma_1^2 \neq \sigma_2^2$$

If H_0 : $\mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$
 (10-15)

is distributed as t-distribution with degrees of freedom given by

$$v = n_1 + n_2 - 2,$$

or more generally

Manhattan plot for Genome-Wide Association Study (GWAS)

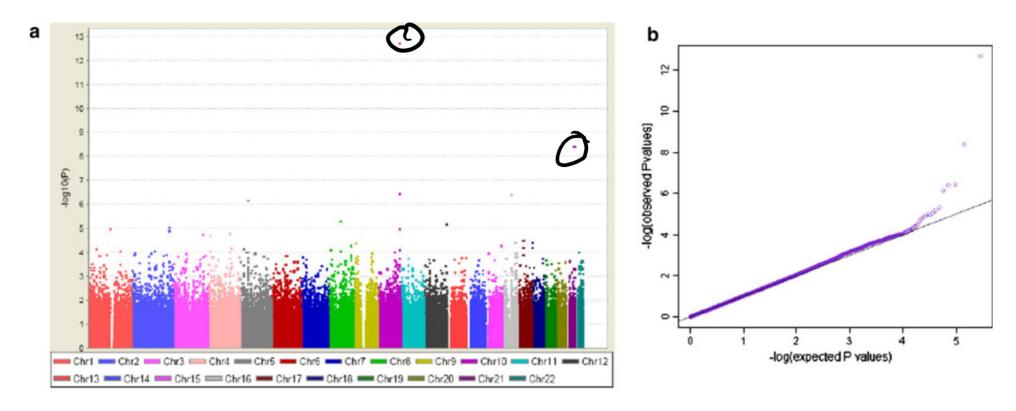


Fig. 1 Genome-wide association results comparing 2,702 cases and 5,726 controls: **a** Manhattan and **b** quantile–quantile plots of $-\log 10$ tranformed *P* values of 285,984 SNPS genotyped

Li J., et al. A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat. 2011;126: 717–727

Multiple null hypotheses: Bonferroni correction

- What if you have m independent null hypotheses?
 Say you have m=25,000 genes in a genome?
- What is the probability that at least one of the null-hypotheses will be shown to be false at significance threshold α_1 ?
- Answer:
 Family-Wise Error Rate or FWER=1-(1- α₁)^m ≈mα₁
- If m=20 and α_1 =0.05, FWER= 0.6415

Carlo Emilio Bonferroni (1892 –1960) Italian mathematician who worked on probability theory.



• If you want to get FWER< α , use $\alpha_1 = \alpha/m$

Chocolate and Cardiovascular Health An article in Nature (2003, Vol. 48, p. 1013) described an

chocolate of In the expe late per day consisted o average bo

Plasma antioxidants from chocolate

Dark chocolate may offer its consumers health benefits the milk variety cannot match.

Is there ev plasma ant

here is some speculation that dietary flavonoids from chocolate, in particular (-)epicatechin, may promote cardiovascular health as a result of direct antioxidant effects or through antithrombotic mechanisms¹⁻³. Here we show that consumption of plain, dark chocolate (Fig. 1) results in an increase in both the total antioxidant capacity and the (-)epicatechin content of blood plasma, but that these effects are markedly reduced when the chocolate is consumed with milk or if milk is incorporated as milk chocolate. Our findings indicate that milk may interfere with the absorption of antioxidants from chocolate in vivo and may therefore negate the potential health benefits that can be derived from eating moderate amounts of dark chocolate.

To determine the antioxidant content of different chocolate varieties, we took dark chocolate and milk chocolate prepared from the same batch of cocoa beans and defatted them twice with *n*-hexane before extracting them with a mixture of water, acetone and acetic acid (70.0:29.8:0.2 by volume). We measured their in vitro total antioxidant capacities using the ferric-reducing antioxidant potential (FRAP) assay4; FRAP

reduced iron per 100 g for dark and milk chocolate, respectively. Volunteers must therefore consume twice as much milk chocolate as dark chocolate to receive a similar intake of antioxidants.

We recruited 12 healthy volunteers (7 women and 5 men with an average age of 32.2 ± 1.0 years (range, 25–35 years). Subjects were non-smokers, had normal blood lipid levels, were taking no drugs or vitamin supplements, and had an average weight of 65.8 ± 3.1 kg (range, 46.0-86.0 kg) and body-mass index of 21.9 ± 0.4 kg m⁻² $(range, 18.6-23.6 \text{ kg m}^{-2})$. On different days, following a crossover experimental design, subjects consumed 100 g dark chocolate, 100 g dark chocolate with 200 ml full-fat milk, or 200 g milk chocolate (containing the equivalent of up to 40 ml milk).

One hour after subjects had ingested the chocolate, or chocolate and milk, we measured the total antioxidant capacity of their plasma by FRAP assay. Plasma antioxidant levels increased significantly after consumption of dark chocolate alone, from $100 \pm 3.5\%$ to $118.4 \pm 3.5\%$ (t-test. P < 0.001), returning to baseline values $(95.4 \pm 3.6\%)$ after 4 h (Fig. 2a). There was



Mauro Serafini*, Rossana Bugianesi*, Giuseppe Maiani*, Silvia Valtuena*, Simone De Santis*, Alan Crozier†

*Antioxidant Research Laboratory, Unit of Human Nutrition, National Institute for Food and Nutrition Research, Via Ardeatina 546. 00178 Rome, Italy

e-mail: serafini@inran.it

†Plant Products and Human Nutrition Group, Graham Kerr Building, Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Figure 1 Stack of benefits? Unlike its milky counterpart, dark chocolate may provide more than just a treat for the tastebuds.

could be due to the formation of secondary bonds between chocolate flavonoids and milk proteins^{6,7}, which would reduce the biological accessibility of the flavonoids and therefore the chocolate's potential antioxidant properties in vivo.

Our findings highlight the possibility

Sweet matlab exercise #1

- Download dark_vs_milk_chocolate_analysis_template.m at the course website. Correct all ?? In the file
- dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];
- milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]
- Use Z-statistics to calculate P-value of the null hypothesis H₀ that milk = dark against H₁ that dark > milk. P_value_z=2*[1-normcdf(|Z|)]
- Repeat using T-statistics. # of degrees of freedom is dof=2*(n-1) P_value_t=2*tcdf(|T|, dof)

Sweet matlab exercise #1

- dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];
- milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]
- x_dark=mean(dark) % sample mean dark chocolate
- x_milk=mean(milk) % sample mean milk chocolate
- s_dark=std(dark) % sample std dark chocolate
- s_milk=std(milk) % sample std milk chocolate
- n=12 % sample size of both dark and milk
- std_xdiff=sqrt(s_dark.^2./2+s_milk.^2./n) % std diff x
- z_stat=(x_dark-x_milk)./std_xdiff % z-statistic
- P_value_z=erfc(z_stat./sqrt(2))./2 % P-value of null true
- % P_value_z=9.9629e-34
- dof=(n-1)+(n-1) % # of degrees of freedom
- P_value_t=tcdf(z_stat,dof,'upper') % P-value of null true
- %P_value_t= 1.8417e-11

