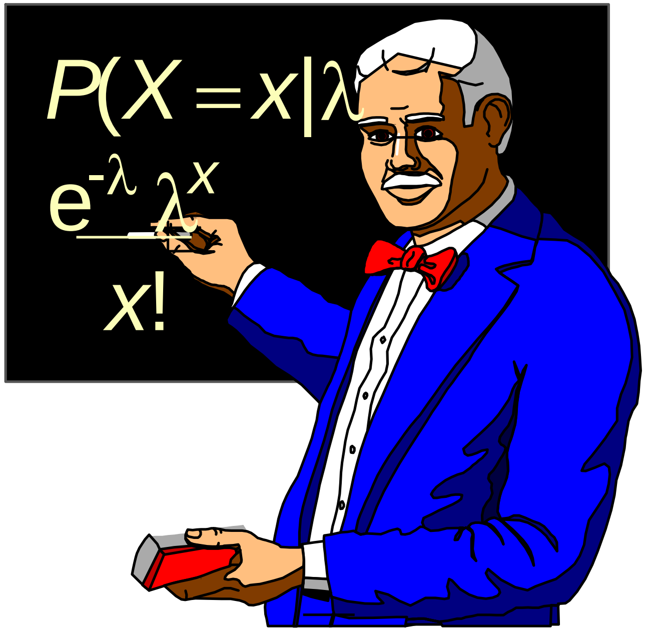
# Phys 443
# Computational Physics

## Hypothesis Testing



# Confidence Levels (CL)

- Suppose an experiment is looking for the *X* particle but observes no candidate events.
- What can we say about the average number of *X* particles expected to have been produced?

  - First, we need to pick a *pd (or pdf)*. Since events are discrete we need a discrete *pd* ⇒ Poisson.
  - Next, how *unlucky* do you want to be? It is common to pick 10% of the time to be *unlucky*.
  - We can now re-state the question as:

"Suppose an experiment finds zero candidate events. What is the 90% CL upper limit on the average number of events (μ) expected assuming a Poisson probability distribution ?"

  - An informal definition of a confidence level (CL):

    CL = 100 x [probability of the event happening by chance]

    The 100 in the above formula allows CL's to be expressed as a percent (%).
  - We can formally write for a continuous probability distribution *p*:

$$\mathrm{CL} = 100 \times prob(x_1 \leq X \leq x_2) = 100 \times \int_{x_1}^{x_2} p(x)dx$$

For a CL we know *p*(x), $x_1$, and $x_2$

# Confidence Intervals

Example: Assume we have a gaussian *pdf* with μ=3 and σ=1. What is the 68% CI ?

We need to solve the following equation:

$$0.68 = \int_a^b G(x,3,1)dx$$

need to solve for *a* and *b*.

Here G(*x*,3,1) is the gaussian *pdf* with μ=3 and σ=1.

There are infinitely many solutions to the above equation.
We (usually) seek the solution that is symmetric about the mean (μ):

$$0.68 = \int_{\mu-c}^{\mu+c} G(x,3,1)dx$$

To solve this problem we either need a probability table, or remember that ≈68% of the area of a gaussian is within ±σ of the mean.

Thus for this problem the 68% CI interval is: [2,4]

# Upper Limits/Lower Limits

- Example: Suppose an experiment observed **no events** of a certain type they were looking for.
- What is the 90% CL upper limit on the expected number of events?

$$\mathrm{CL} = 0.90 = \sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^{n}}{n!}$$

$$1 - \mathrm{CL} = 0.10 = 1 - \sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^{n}}{n!} = \sum_{n=0} \frac{e^{-\lambda}\lambda^{n}}{n!} = e^{-\lambda}$$

$$\lambda = 2.3$$

So, if μ=2.3 then 10% of the time we should expect to find 0 candidates. There was nothing wrong with our experiment. We were just unlucky.

# Confidence Levels (CL)

Example: A cosmic ray experiment with effective area=$10^3$ km$^2$ looks for events with energies >$10^{20}$ eV and after one year has no candidate events. We can calculate a 90% UL on the flux of these high energy events:

$$\boxed{\text{Flux} < 2.3\text{x}10^{-3}/\text{km}^2/\text{year @ 90\% CL.}}$$

# Poisson Upper Limits

Example: Suppose an experiment finds one candidate event. What is the 95% CL upper limit on the average number of events (μ) ?

$$1 - CL = 1 - \sum_{n=2}^{\infty} \frac{e^{-\mu}\mu^{n}}{n!} = \sum_{n=0}^{1} \frac{e^{-\mu}\mu^{n}}{n!} = e^{-\mu} + \mu e^{-\mu} \Rightarrow \mu = 4.74$$

The 5% includes 1 AND 0 events.

Here we are saying that we would get 2 *or more* events 95% of the time if μ=4.74.

# Poisson Upper Limits

- Things get much more interesting when we have background in our data sample! We measure N events & we predict B background events. The number of our signal events, S, is: S=N-B

## Three interesting situations can arise:

I) How should we handle the case where B ≥ N ?

Usually B is calculated without knowledge of the value of N.
Since B and N are obtained independently there is nothing that guarantees that N>B

II) Even if N > B a sloppy background prediction can lead to a better (smaller) UL than a careful background prediction!

For fixed N, larger B ⇨means smaller value of S ⇨ smaller UL.

III) More background is better than less background?

# Confidence Regions & MLM

- Often we have a problem that involves two or more parameters. In these cases it makes sense to **define confidence regions rather than an interval**.

- Consider the case where we are doing a MLM fit to two variables α, β.
- Previously we have seen that for large samples the Likelihood function becomes “Gaussian”:

$$\ln L(\alpha) = \ln L_{\max} - \frac{1}{2}\frac{(\alpha-\alpha^*)^2}{\sigma_\alpha^2}$$

Consider the case of two correlated variables α, β:

$$\ln L(\alpha,\beta) = \ln L_{\max} - \frac{1}{2}\frac{1}{(1-\rho^2)}\left[\frac{(\alpha-\alpha^*)^2}{\sigma_\alpha^2} + \frac{(\beta-\beta^*)^2}{\sigma_\beta^2} - 2\rho\frac{(\alpha-\alpha^*)(\beta-\beta^*)}{\sigma_\alpha\sigma_\beta}\right] \quad with \quad \rho = \frac{\sigma_{\alpha\beta}}{\sigma_\alpha\sigma_\beta}$$

# Confidence Regions & MLM

The contours of constant likelihood are given by:

$$\ln L(\alpha,\beta) = \ln L_{\max} - \frac{1}{2}Q \Rightarrow L(\alpha,\beta) = L_{\max}\, e^{-\frac{1}{2}Q}$$

$$Q \equiv \frac{1}{(1-\rho^2)}\left[\frac{(\alpha-\alpha^*)^2}{\sigma_\alpha^2} + \frac{(\beta-\beta^*)^2}{\sigma_\beta^2} - 2\rho\frac{(\alpha-\alpha^*)(\beta-\beta^*)}{\sigma_\alpha\sigma_\beta}\right]$$

We can re-write Q in matrix form, with V$^{-1}$ the inverse of the error matrix:

$$Q \equiv \frac{1}{(1-\rho^2)}\begin{pmatrix}\alpha-\alpha^*\\ \beta-\beta^*\end{pmatrix}^T \begin{pmatrix}1/\sigma_\alpha^2 & -\rho/\sigma_\alpha\sigma_\beta\\ -\rho/\sigma_\alpha\sigma_\beta & 1/\sigma_\beta^2\end{pmatrix}\begin{pmatrix}\alpha-\alpha^*\\ \beta-\beta^*\end{pmatrix} = \overline{\alpha}^T V^{-1}\overline{\alpha}$$

We can generalize this to *n* parameters with V the *nxn* error matrix and α an *n*-dimensional vector

# Confidence Regions & MLM

- The variable Q is described by a $\chi^2$ pdf with #dof= # of parameters
  - For the case where the parameters have gaussian pdfs this is exact
  - For the non-gaussian case this is true in the limit of a large data sample

Note: we can re-write Q in the expected form of a $\chi^2$ variable by transforming from correlated to uncorrelated variables.

For the 2D case the transformation is a rotation:

$$\begin{pmatrix} x-x^* \\ y-y^* \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \alpha-\alpha^* \\ \beta-\beta^* \end{pmatrix} \quad \text{with} \quad \tan 2\theta = \frac{2\rho\sigma_\alpha\sigma_\beta}{\sigma_\alpha^2-\sigma_\beta^2}$$

$$p(\chi^2,n) = \frac{1}{2^{n/2}\Gamma(n/2)}[\chi^2]^{n/2-1}e^{-\chi^2/2} \Rightarrow p(\chi^2,2) = \frac{1}{2}e^{-Q/2}$$

**Since we know the pdf for Q we can calculate a confidence level for a fixed value of Q. The case of 2 variables is easy since the $\chi^2$ pdf is just:**

$$\left[\frac{(\alpha-\alpha^*)^2}{\sigma_\alpha^2} + \frac{(\beta-\beta^*)^2}{\sigma_\beta^2} - 2\rho\frac{(\alpha-\alpha^*)(\beta-\beta^*)}{\sigma_\alpha\sigma_\beta}\right] \Rightarrow \frac{(x-x^*)^2}{\sigma_x^2} + \frac{(y-y^*)^2}{\sigma_y^2}$$

# Confidence Regions & MLM

We can calculate the confidence level for the region bounded by **a fixed value of Q**:

$$CL = P(Q \le Q_0) = {}^1\!/_2 \int_0^{Q_0} e^{-\frac{1}{2}Q}\, dQ = 1 - e^{-\frac{1}{2}Q_0}$$

| $Q_0$ | CL(%) |
|---|---|
| 1 | ≈39 |
| 2.3 | ≈68 |
| 4.6 | ≈90 |
| 6.2 | ≈95 |
| 9.2 | ≈99 |

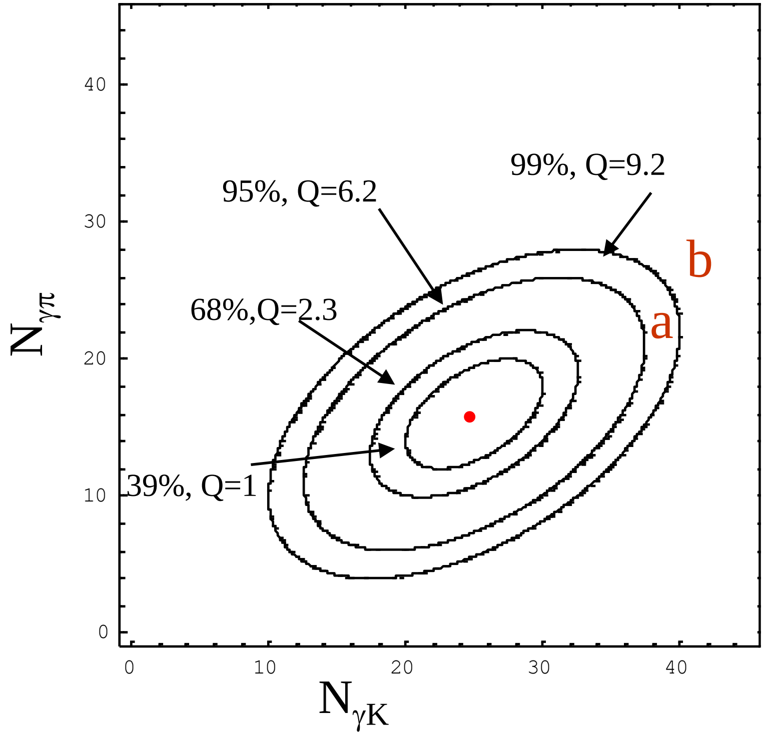
# Confidence Regions & MLM

Example: Suppose in an experiment a maximum likelihood analysis to search for B→$\gamma\pi^-$ and B→$\gamma K^-$ events. The results of the MLM fit are:

$N_{\gamma\pi}$ =16±4, $N_{\gamma K}$ =25±5, ρ=0.5

$N_{\gamma\pi}$ and $N_{\gamma K}$ are highly correlated, since at high momentum (>2GeV) it has a hard to separate π's and K's.

$$Q = \frac{1}{(1-.5^2)}\left[\frac{(N_{\gamma K}-25)^2}{5^2}+\frac{(N_{\gamma\pi}-16)^2}{4^2}-2(0.5)\frac{(N_{\gamma K}-25)(N_{\gamma\pi}-16)}{5\times 4}\right]$$



The contours of constant probability are given by:

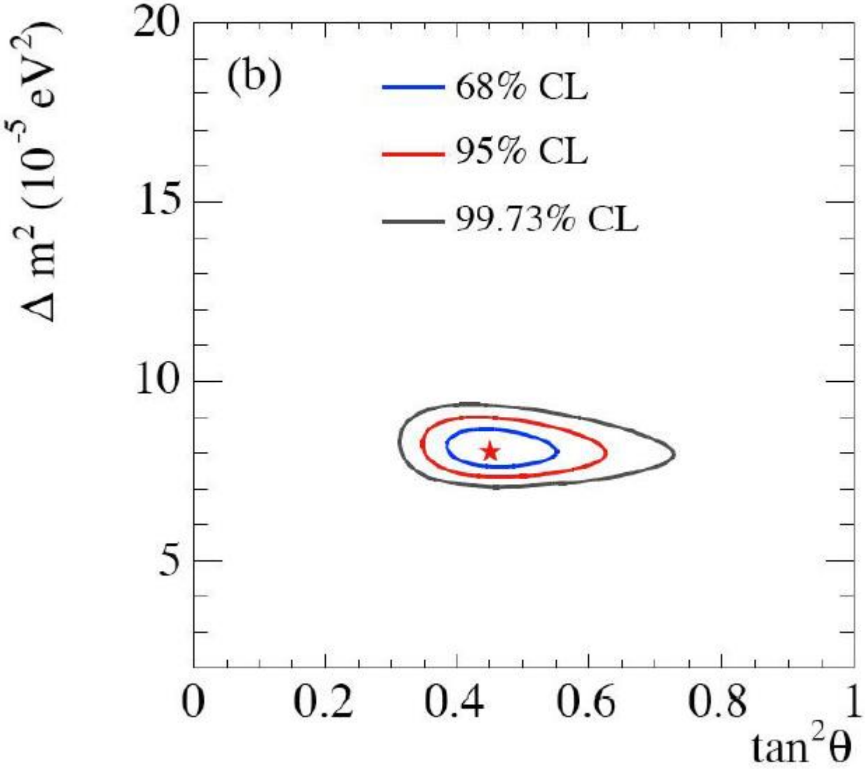
Point "a" is excluded at the 95%CL

Point "b" is excluded at the 99%CL

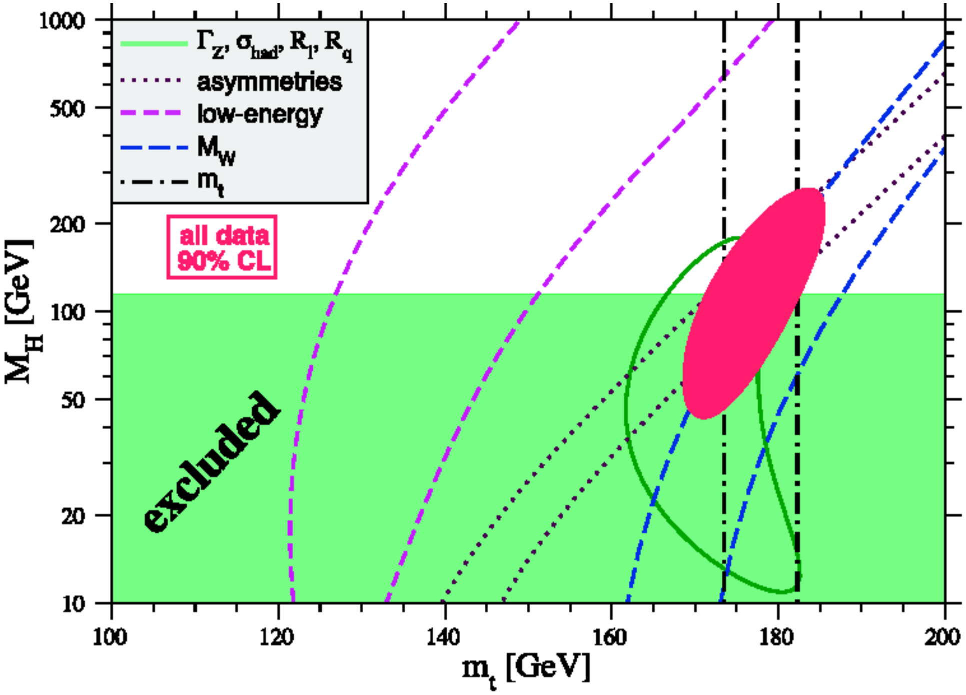
| $Q_0$ | CL(%) |
|---|---|
| 1 | ≈39 |
| 2.3 | ≈68 |
| 4.6 | ≈90 |
| 6.2 | ≈95 |
| 9.2 | ≈99 |

# Examples

solar neutrino oscillations experiments



mass of Higgs Vs mass of top quark



Both examples show allowed regions at various confidence levels.

# Hypothesis Testing

- **Suppose that there is a hypothesis/theory that you don’t want to reject unless there is evidence that it is false**.
  - If the data you collect contains evidence that the hypothesis is false, you will reject it.
  - If not, you won’t.

- A classic example is a court trial: it is assumed that you are not guilty unless proven otherwise. Another example is testing a new drug: being conservative, we assume that the new drug is ineffective (or even detrimental) unless proven otherwise.

# Hypothesis Testing

- Our mathematical framework will be the following:
  - we have data (which in this course we will always assume to be $X_1,X_2, \dots ,X_n$ coming from a model with PDF $f_\theta(x)$, where $\theta$ is unknown. The hypothesis that
  - **we assume "true" unless proven otherwise is called the null hypothesis ($H_0$), which is tested against the alternative hypothesis ($H_1$).**
  - The alternative hypothesis is not accepted unless there is evidence that supports it.
- In our setup, $H_0$ will correspond to assuming that $\theta= \theta_0$ for a fixed value of $\theta_0$(sometimes we might consider $H_0$s of the type $\theta\leq \theta_0$ or $\theta\geq \theta_0$ , which turn out to be "equivalent" in some sense to working with $H_0 : \theta= \theta_0$ ).

# Hypothesis  Testing

To answer this question in a quantitative way, one has to
– pose a hypothesis
– define accept and rejection conditions & perform the test

There are two types of hypotheses tests: parametric and non-parametric

Parametric: compares the values of parameters (e.g. does the mass of proton = mass of electron ?)

Non-parametric: deals with the shape of a distribution  (e.g. is angular distribution consistent with being flat?)

# Example I

Hypothesis testing:
We have a certain theory and want to know whether it agrees or disagrees with our measurement

Example:
A theory predicts that BR(Higgs$\rightarrow\mu^+\mu^-$) = $2\text{x}10^{-5}$ and you measure $(4\pm2)\ \text{x}10^{-5}$ .

The hypothesis we want to test is "are experiment and theory consistent?"

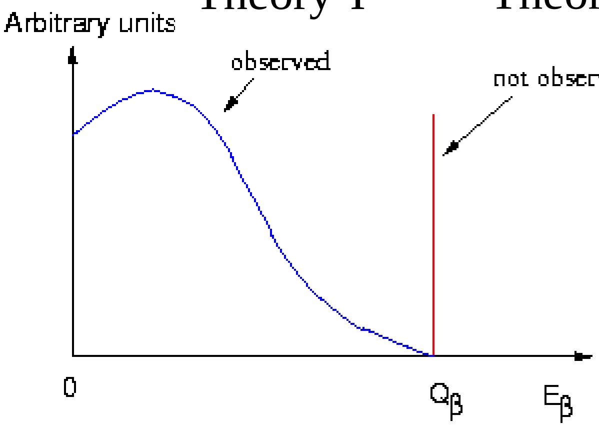
Example:
An experiment measures the $\Lambda_c$ lifetime to be (180 ± 7)fs while another experiment measures (198 ± 7) fs. The hypothesis we want to test is "are the lifetime results from these experiments are consistent?"

# Example II

- Consider the case of beta decay. Suppose we have two theories that both predict the energy spectrum of the electron emitted in the decay of the neutron. Here a parametric test might not be able to distinguish between the two theories since both theories **might predict the same average energy of the emitted electron**.
- However a non-parametric test would be able to distinguish between the two theories as the shape of the energy spectrum differs for each theory.

Theory-I

Theory-II



$$ {}^{A}_{Z}X \rightarrow {}^{A}_{Z+1}Y + e^{-} $$

$$ {}^{A}_{Z}X \rightarrow {}^{A}_{Z+1}Y + e^{-} + \nu $$

# Hypothesis Testing

We start with a null hypothesis ($H_0$) that represents the status quo. We develop an alternative hypothesis ($H_A$) that represents our research question (what we're testing for).We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods.

- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis.
- If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

# Bayesian hypothesis testing

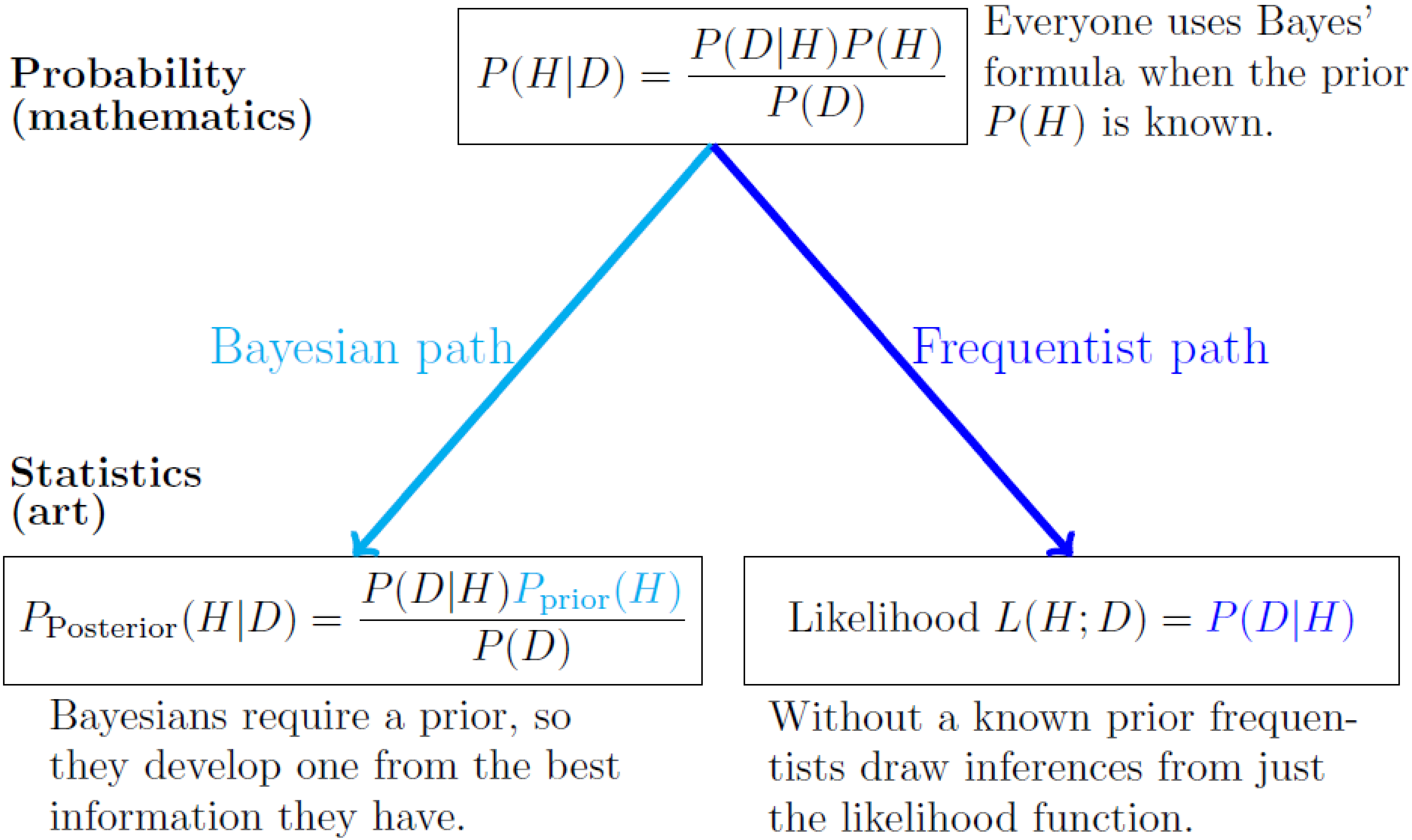
As usual, Bayesian analysis is far more straightforward than the frequentist version: in Bayesian language, all problems are hypothesis tests!

$$P(H|D,I)=\frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

- Bayesian hypothesis testing requires you to explicitly specify the alternative hypotheses. This comes about when calculating
  P(D|I)=P(D|H1,I)+P(D|H2,I)+P(D|H3,I) ...

- **Hypothesis testing is more sensitive to priors than parameter estimation**. For example, hypothesis testing may involve

# hypothesis testing



**Probability (mathematics)**

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Everyone uses Bayes' formula when the prior $P(H)$ is known.

Bayesian path

Frequentist path

**Statistics (art)**

$$P_{\text{Posterior}}(H|D) = \frac{P(D|H)P_{\text{prior}}(H)}{P(D)}$$

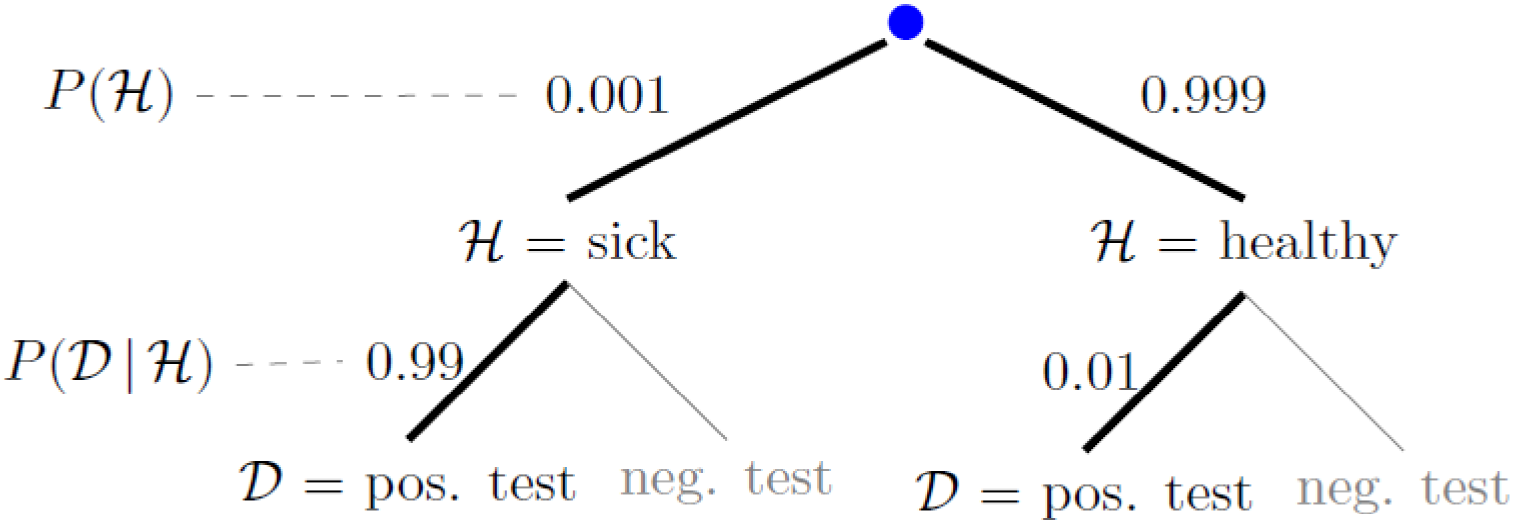
Bayesians require a prior, so they develop one from the best information they have.

Likelihood $L(H; D) = P(D|H)$

Without a known prior frequentists draw inferences from just the likelihood function.

# hypothesis testing

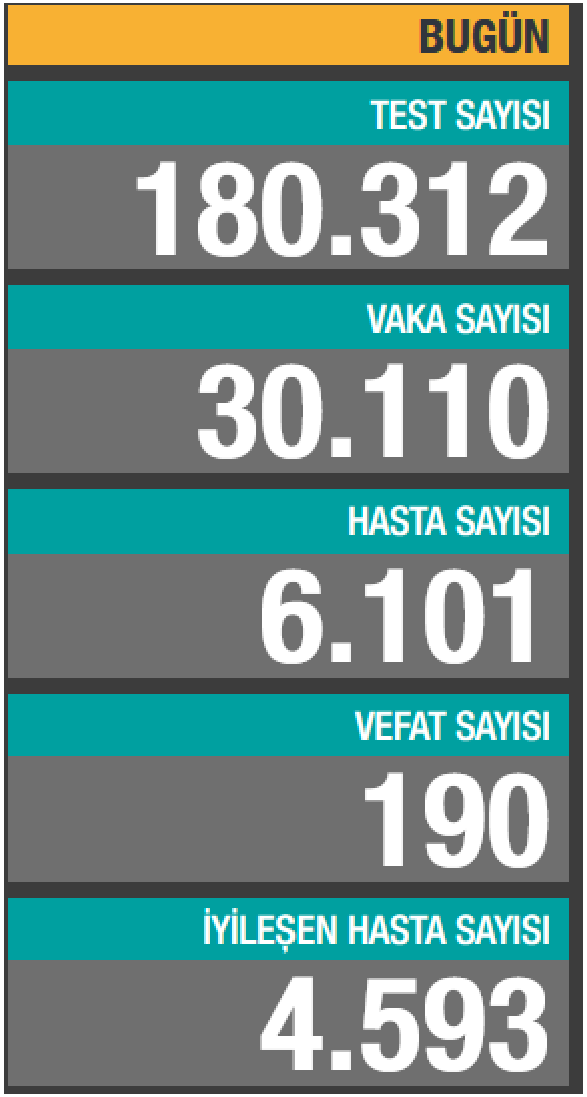
The test is positive. Are you sick?



The prior is known so we can use Bayes' Theorem.

$$P(\text{sick}\,|\,\text{pos. test}) = \frac{0.001 \cdot 0.99}{0.001 \cdot 0.99 + 0.999 \cdot 0.01} \approx 0.1$$

# hypothesis testing: COVİD-19



BUGÜN

TEST SAYISI

180.312

VAKA SAYISI

30.110

HASTA SAYISI

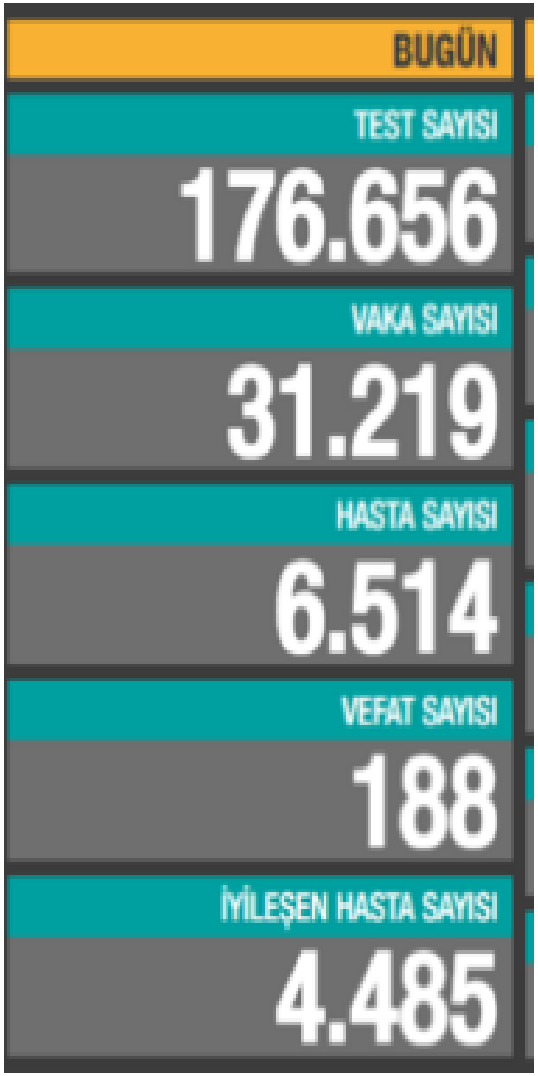
6.101

VEFAT SAYISI

190

İYİLEŞEN HASTA SAYISI

4.593



BUGÜN

TEST SAYISI

176.656

VAKA SAYISI

31.219

HASTA SAYISI

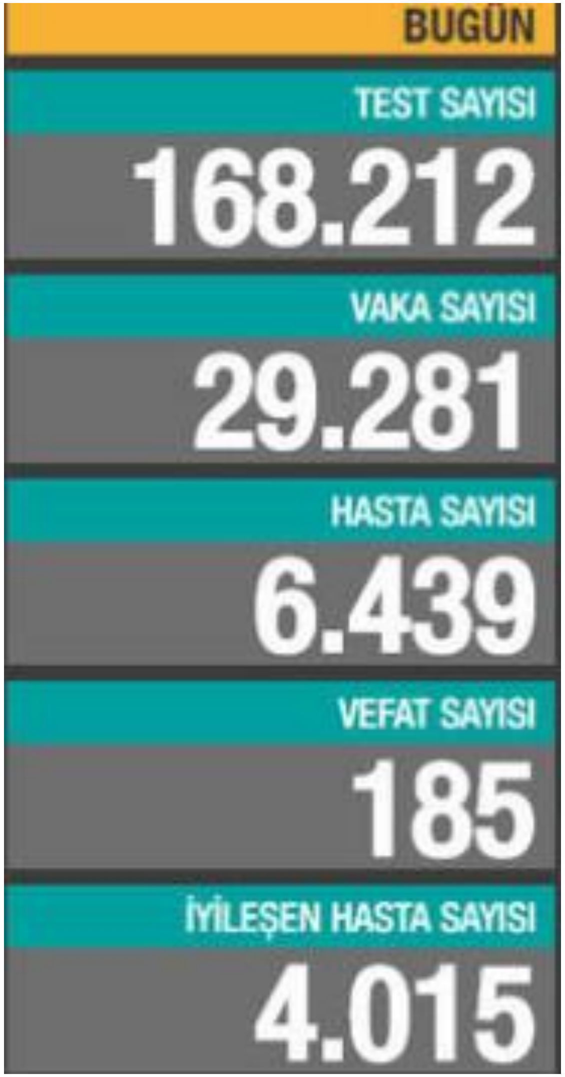
6.514

VEFAT SAYISI

188

İYİLEŞEN HASTA SAYISI

4.485



BUGÜN

TEST SAYISI

168.212

VAKA SAYISI

29.281

HASTA SAYISI

6.439

VEFAT SAYISI

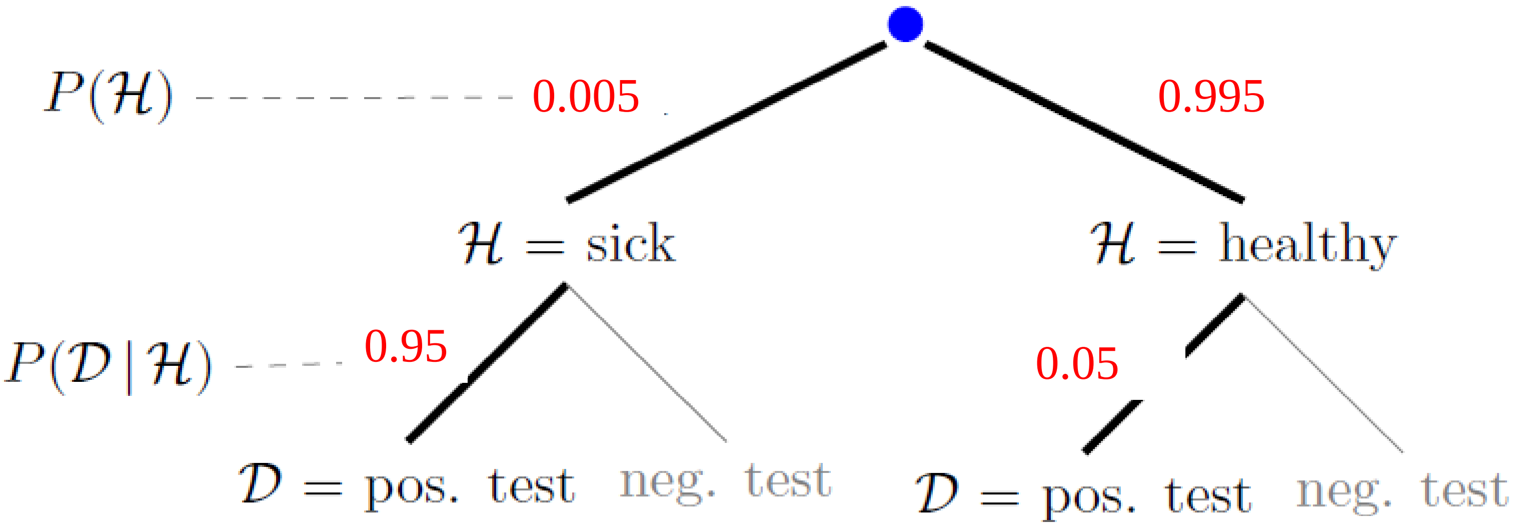
185

İYİLEŞEN HASTA SAYISI

4.015

# hypothesis testing: COVİD19

The test is positive. Are you sick?

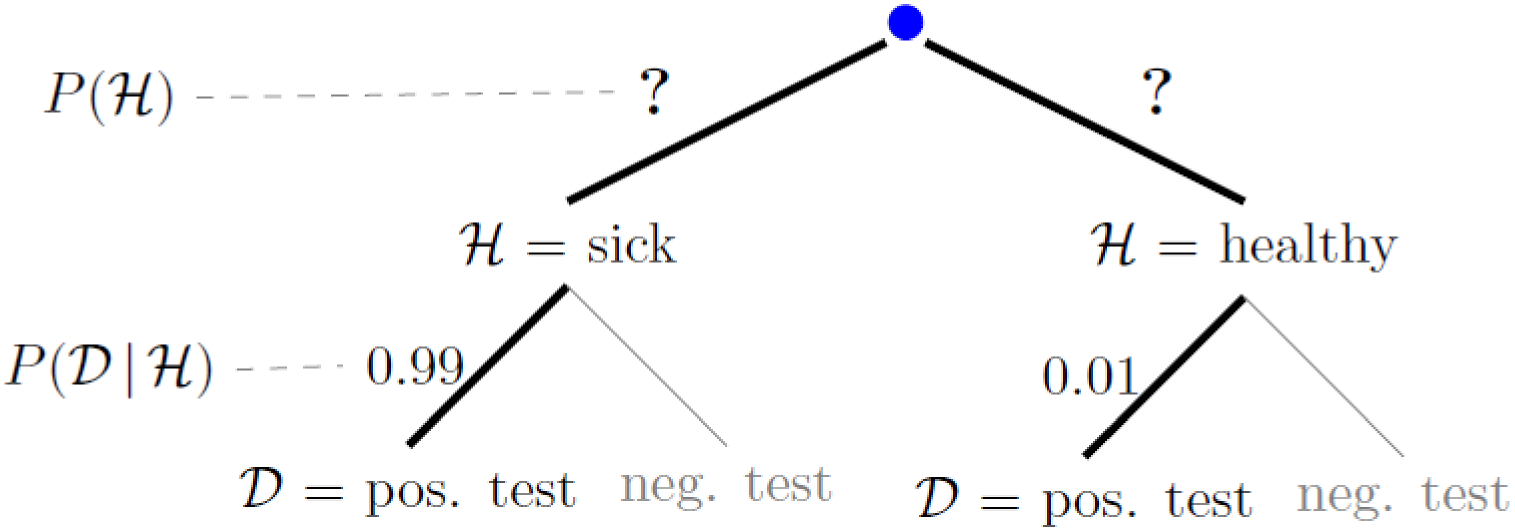


The prior is known so we can use Bayes' Theorem.

$$P(\text{sick}\mid\text{pos. test}) = \frac{0.005 \mathrm{~x~} 0.95}{0.005 \mathrm{~x~} 0.95 + 0.995 \mathrm{~x~} 0.05} \simeq 0.087$$

# hypothesis testing

The test is positive. Are you sick?



$P(\mathcal{H})$
?
?
$\mathcal{H} = \text{sick}$
$\mathcal{H} = \text{healthy}$
$P(\mathcal{D} \,|\, \mathcal{H})$
0.99
0.01
$\mathcal{D} = \text{pos. test}$
neg. test
$\mathcal{D} = \text{pos. test}$
neg. test

The prior is not known.
Bayesian: use a subjective prior $P(\mathcal{H})$ and Bayes' Theorem.

Frequentist: the likelihood is all we can use: $P(\mathcal{D} \,|\, \mathcal{H})$

# Classical frequentist testing: Type I errors

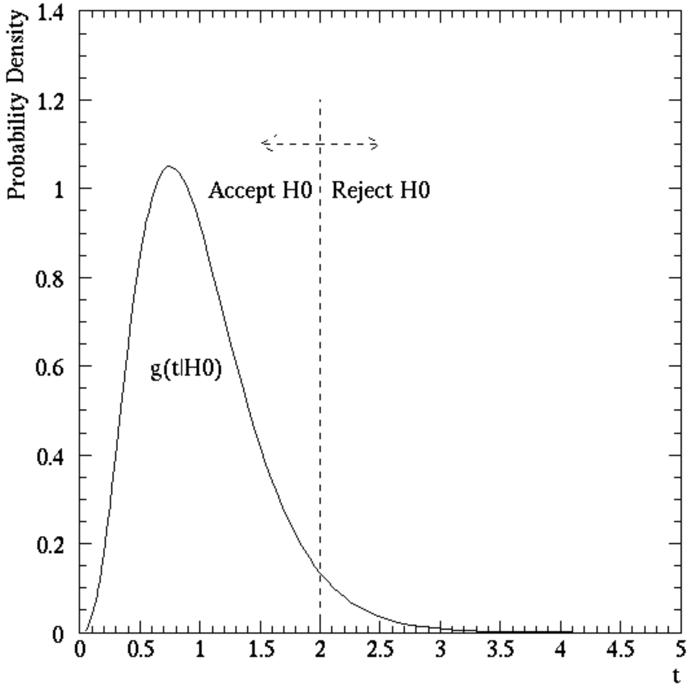
In frequentist hypothesis testing, we construct a test statistic from the measured data, and use the value of that statistic to decide whether to accept or reject the hypothesis.

| | Actual Truth of $H_0$ | |
|---|---|---|
| **Decision** | **$H_0$ is true** | **$H_0$ is false** |
| Do not reject $H_0$ | Correct Decision | **Type II Error** |
| Reject $H_0$ | **Type I Error** | Correct Decision |

# Classical frequentist testing: Type I errors

- **Type 1 error** can only be made if the null hypothesis is actually true.

Type I Error: You reject a true hypothesis



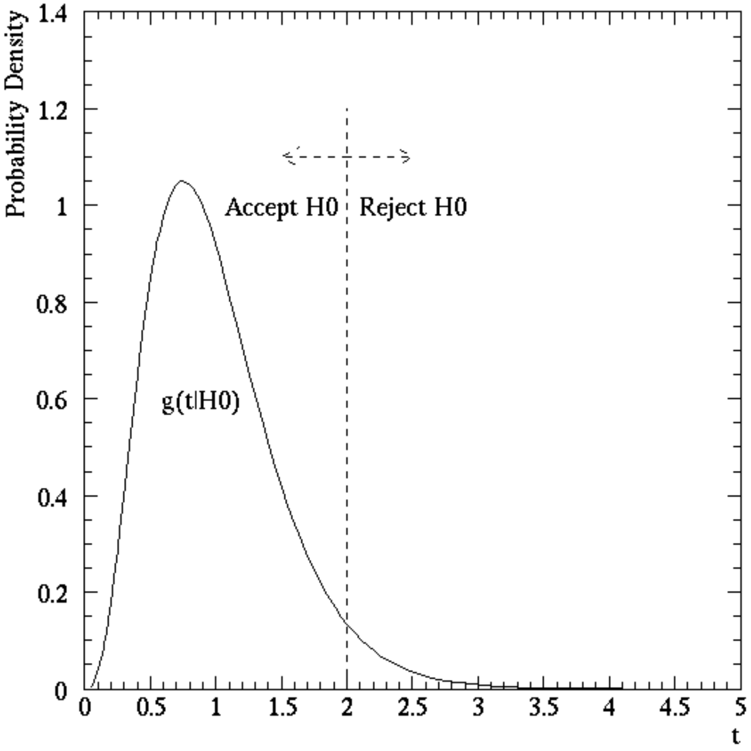
# Classical frequentist testing: Type II errors

- **Type 2 error** can only be made if the alternative hypothesis is actually true.

Type II error: We accept the hypothesis $H_0$ even though it is false, and instead $H_1$ is really true. Probability = area on tail of g(t|H1)= β

$$\beta = \int_{-\infty}^{t_{cut}} dt\, g(t|H1)$$

Often you choose what probability you're willing to accept for Type I errors (falsely rejecting your hypothesis), and then choose your cut region to minimize β. You have to specify the alternate hypothesis if you want to determine β.



# Hypothesis Testing

**Some Notation for Hypothesis Tests**

The null hypothesis is denoted by $H_0$, and the alternative hypothesis is denoted by $H_1$ or $H_A$

"alpha" = $\alpha$ = desired probability of making a type 1 error when $H_0$ is true; we reject $H_0$ if *p*-value $\leq \alpha$.

"beta" = $\beta$ = probability of making a type 2 error when $H_1$ is true; power = $1 - \beta$

**Steps for Testing the Mean of a Single Population**

Denote the population mean by $\mu$ and the sample mean and standard deviation by $\overline{X}$ and *s*, respectively.

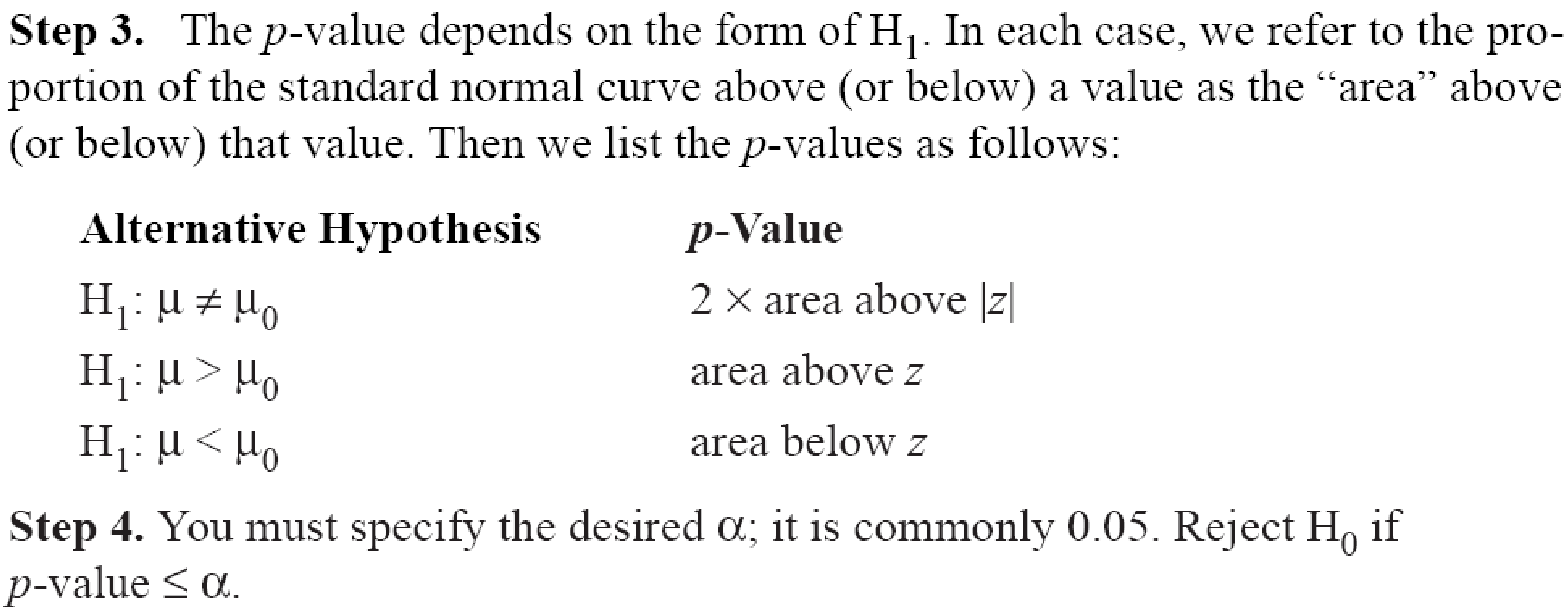
**Step 1.** $H_0$: $\mu = \mu_0$, where $\mu_0$ is the *chance* or *status quo* value.

$H_1$: $\mu \neq \mu_0$ for a two-sided test; $H_1$: $\mu < \mu_0$ or $H_1$: $\mu > \mu_0$ for a one-sided test, with the direction determined by the research hypothesis of interest.

**Step 2.** This test statistic applies only if the sample is large. The test statistic is

$$z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

# Hypothesis Testing



**Step 3.** The $p$-value depends on the form of $H_1$. In each case, we refer to the proportion of the standard normal curve above (or below) a value as the "area" above (or below) that value. Then we list the $p$-values as follows:

| **Alternative Hypothesis** | ***p*-Value** |
|---|---|
| $H_1$: $\mu \neq \mu_0$ | 2 × area above $\lvert z \rvert$ |
| $H_1$: $\mu > \mu_0$ | area above $z$ |
| $H_1$: $\mu < \mu_0$ | area below $z$ |

**Step 4.** You must specify the desired $\alpha$; it is commonly 0.05. Reject $H_0$ if $p$-value $\leq \alpha$.

# Hypothesis Testing

**Steps for Testing a Proportion for a Single Population**

Steps 1, 3, and 4 are the same, except replace μ with the population proportion *p* and $\mu_0$ with the hypothesized proportion $p_0$. The test statistic (step 2) is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

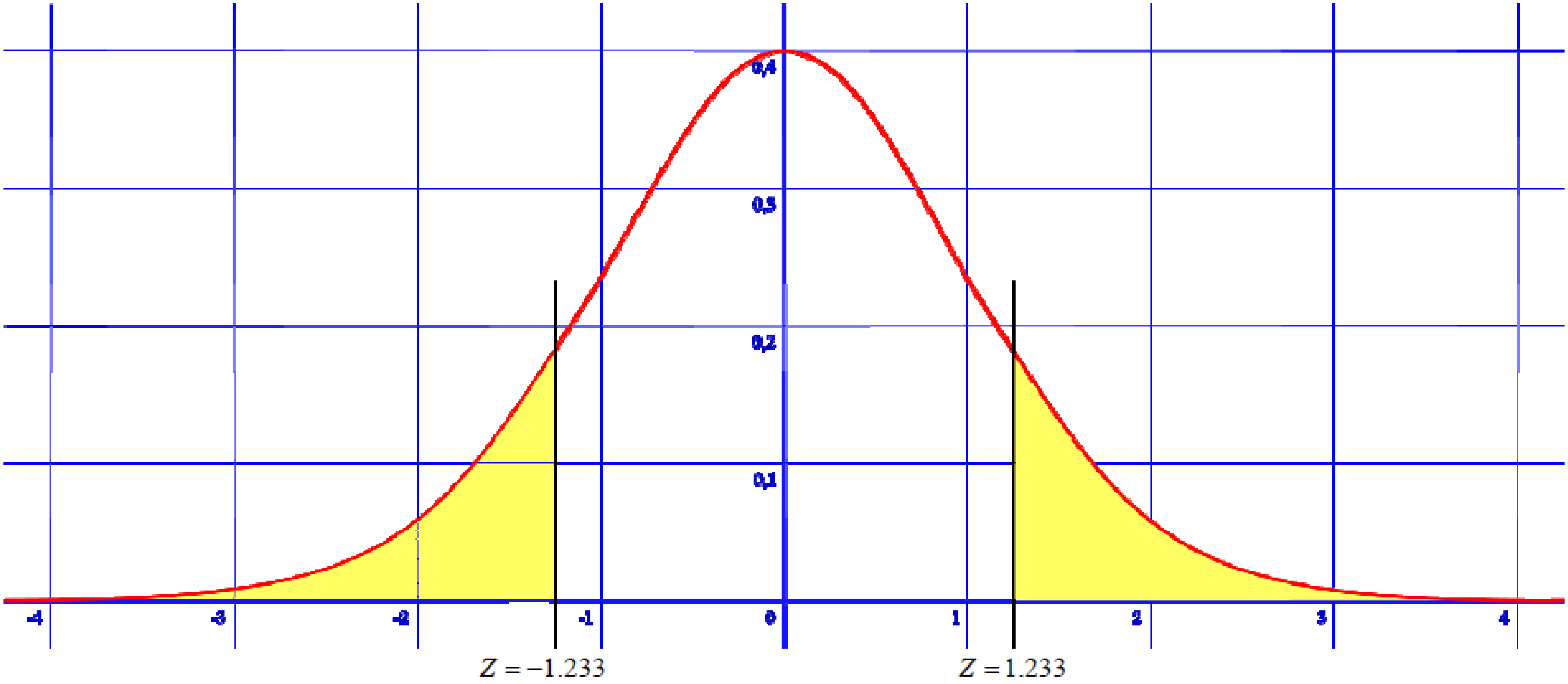
**Steps for Testing for Equality of Two Population Means Using Large Independent Samples**

Steps 1, 3, and 4 are the same, except replace μ with $(\mu_1 - \mu_2)$ and $\mu_0$ with 0.

Use previous notation for sample sizes, means, and standard deviations; the test statistic (step 2) is:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}}}$$

# P-value



$$\text{p-value} \;=\; \mathrm{Prob}\left(|Z| \geq |Z_{\text{obs}}|\right) \;=\; 1 - \int_{-Z_{\text{obs}}}^{Z_{\text{obs}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} z^2\right) dz$$

# Making a decision - p-values:example

In 2010 the average GPA of students at METU was 3.37. Last semester METU students in a Phys112 class were surveyed and ask for their current GPA. This survey had 147 respondents and yielded an average GPA of 3.56 with a standard deviation of 0.31.

•Assuming that this sample is random and representative of all METU students, do these data provide convincing evidence that the average **GPA of METU students has changed over the last decade?**

# Setting the hypotheses

- The parameter of interest is the average GPA of current METU students.
- There may be two explanations why our sample mean is higher than the average GPA from 2010.
  - The true population mean has changed.
  - The true population mean remained at 3.37, the difference between the true population mean and the sample mean is simply due to natural sampling variability.

- We start with the assumption that nothing has changed.

  $H_0 : \ \mu = 3.37$

- We test the claim that average GPA has changed.

  $H_A : \ \mu \neq 3.37$

# Calculating the p-value

p-value: probability of observing data at least as favorable to $H_A$ as our current data set (a sample mean greater than 3.56 or less than 3.18), if in fact $H_0$ was true (the true population mean was 3.37).

$$\begin{aligned}
&P(\bar{x} > 3.56 \text{ or } \bar{x} < 3.18 \mid \mu = 3.37) \\
&\quad = P(\bar{x} > 3.56 \mid \mu = 3.37) + P(\bar{x} < 3.18 \mid \mu = 3.37) \\
&\quad = P\left( \mathrm{z} > \frac{3.56 - 3.37}{0.31/\sqrt{147}} \right) + P\left( \mathrm{z} < \frac{3.18 - 3.37}{0.31/\sqrt{147}} \right) \\
&\quad = P(\mathrm{z} > 7.43) + P(\mathrm{z} < -7.43) \\
&\quad = 10^{-13} \approx 0
\end{aligned}$$

# Inference/draw a conclusion

p-value: probability of observing data at least as favorable to $H_A$ as our current data set (a sample mean greater than 3.56 or less than 3.18), if in fact $H_0$ was true (the true population mean was 3.37).

$p$ –value $\approx 10^{-13}$

- If the true average GPA METU students applied to is 3.37, there is approximately a $10^{-11}$ % chance of observing a random sample of 147 METU students with an average GPA of 3.56.

  - This is a very low probability for us to think that a sample mean of 3.56 GPA is likely to happen simply by chance.
  - Since p-value is low (lower than 5%) we reject $H_0$.
  - The data provide convincing evidence that METU students average GPA has changed since 2010.

- There is significant evidence that the difference between the null value of a 3.37 GPA and observed sample mean of 3.56 GPA is not due to chance or sampling variability.

# P-value: Example

- A sample of 51 Panasonic AAA batteries gave a sample mean zinc mass of 2.06 g and a sample standard deviation of 0.141 g.

- Does this data provide compelling evidence for concluding that the population mean zinc mass exceeds 2.0g?

- With μ denoting the true average zinc mass for such batteries, the relevant hypotheses are $H_0$: $\mu = 2.0$ versus $H_A$: $\mu > 2.0$.

- The sample size is large enough so that a z test can be used without making any specific assumption about the shape of the population distribution.

# P-value: Example

- The test statistic value is

$$z = \frac{\bar{x} - 2.0}{s/\sqrt{n}} = \frac{2.06 - 2.0}{.141/\sqrt{51}} = 3.04$$

- Now we must decide which values of z are at least as contradictory to $H_0$.

  - Let’s first consider an easier task:
    Which values of x are at least as contradictory to the null hypothesis as 2.06, the mean of the observations in our sample?

# P-value: Example

- Because > appears in $H_A$, it should be clear that 2.10 is at least as contradictory to $H_0$ as is 2.06, and so in fact is any x value that exceeds 2.06

- But an $\overline{x}$ value that exceeds 2.06 corresponds to a value of z that exceeds 3.04. Thus the P-value is

$$\text{P-value} = P(Z \geq 3.04 \text{ when } \mu = 2.0)$$

- Since the test statistic Z was created by subtracting the null value 2.0 in the numerator, when $\mu = 2.0$—i.e., when $H_0$ is true—Z has approximately a standard normal distribution.

# P-value: Example

As a consequence,

$P$-value = $P(Z \geq 3.04 \text{ when } \mu = 2.0)$

$\approx$ area under the $z$ curve to the right of 3.04

$= 1 - \Phi(3.04)$

$= .0012$

# P-value: Example

We will shortly illustrate how to determine the P-value for any z or t test—i.e., any test where the reference distribution is the standard normal distribution (and z curve) or some t distribution (and corresponding t curve).

For the moment, though, let’s focus on reaching a conclusion once the P-value is available.

Because it is a probability, the P-value must be between 0 and 1.

# Hypothesis Testing

## Some Notation for Hypothesis Tests

The null hypothesis is denoted by $H_0$, and the alternative hypothesis is denoted by $H_1$ or $H_a$.

"alpha" = $\alpha$ = desired probability of making a type 1 error when $H_0$ is true; we reject $H_0$ if *p*-value $\leq \alpha$.

"beta" = $\beta$ = probability of making a type 2 error when $H_1$ is true; power = $1 - \beta$

## Steps for Testing the Mean of a Single Population

Denote the population mean by $\mu$ and the sample mean and standard deviation by $\overline{X}$ and *s*, respectively.

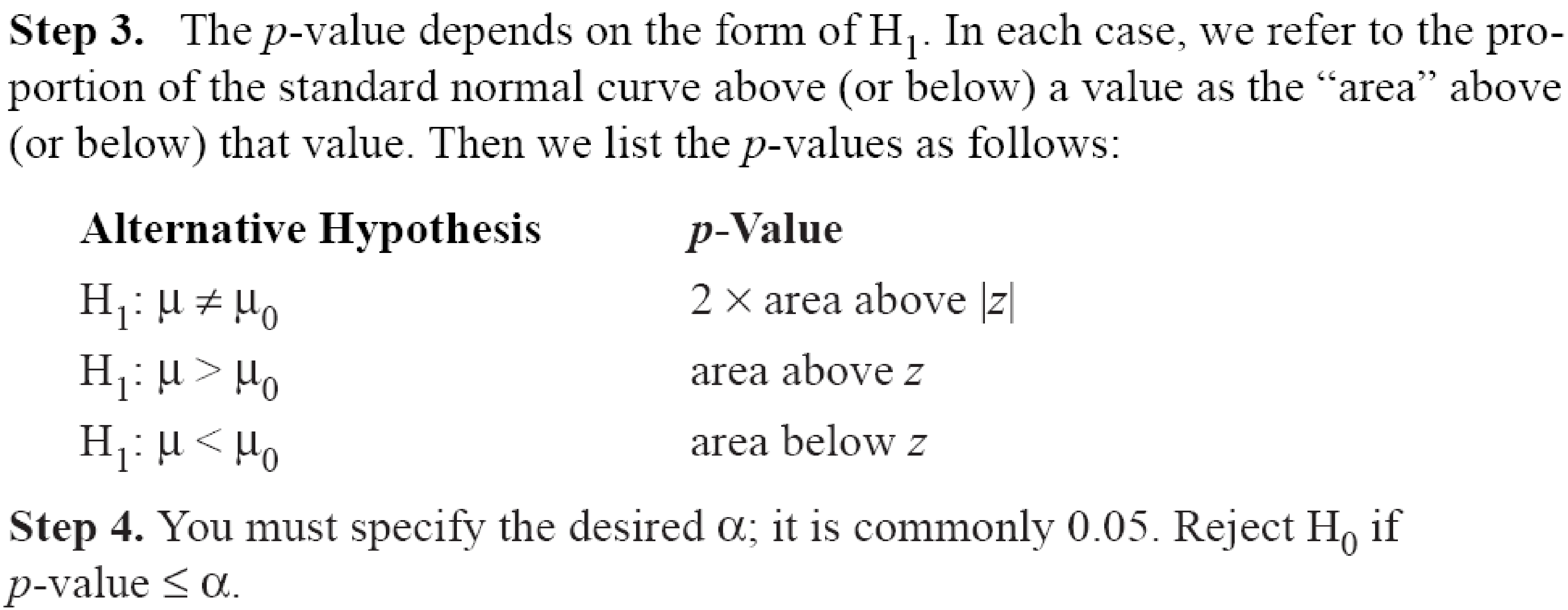
**Step 1.** $H_0$: $\mu = \mu_0$, where $\mu_0$ is the *chance* or *status quo* value.

$H_1$: $\mu \neq \mu_0$ for a two-sided test; $H_1$: $\mu < \mu_0$ or $H_1$: $\mu > \mu_0$ for a one-sided test, with the direction determined by the research hypothesis of interest.

**Step 2.** This test statistic applies only if the sample is large. The test statistic is

$$z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

# Hypothesis Testing

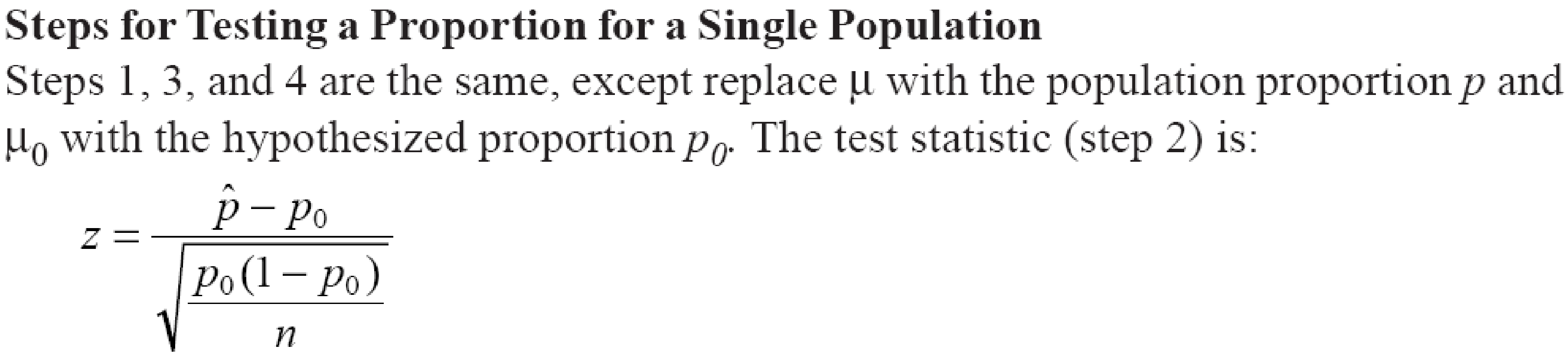


**Step 3.** The $p$-value depends on the form of $H_1$. In each case, we refer to the proportion of the standard normal curve above (or below) a value as the "area" above (or below) that value. Then we list the $p$-values as follows:

| **Alternative Hypothesis** | ***p*-Value** |
|---|---|
| $H_1: \mu \neq \mu_0$ | $2 \times$ area above $\lvert z \rvert$ |
| $H_1: \mu > \mu_0$ | area above $z$ |
| $H_1: \mu < \mu_0$ | area below $z$ |

**Step 4.** You must specify the desired $\alpha$; it is commonly 0.05. Reject $H_0$ if $p$-value $\leq \alpha$.

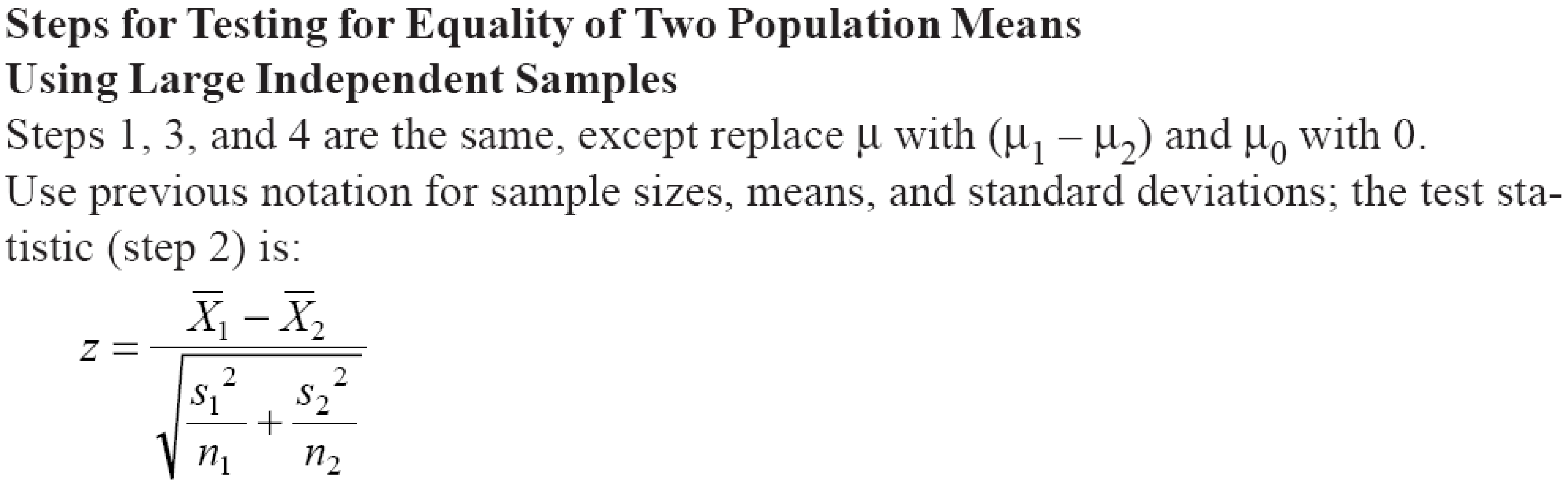
# Hypothesis Testing



**Steps for Testing a Proportion for a Single Population**

Steps 1, 3, and 4 are the same, except replace μ with the population proportion $p$ and $\mu_0$ with the hypothesized proportion $p_0$. The test statistic (step 2) is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$



**Steps for Testing for Equality of Two Population Means Using Large Independent Samples**

Steps 1, 3, and 4 are the same, except replace μ with $(\mu_1 - \mu_2)$ and $\mu_0$ with 0.

Use previous notation for sample sizes, means, and standard deviations; the test statistic (step 2) is:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1{}^2}{n_1} + \dfrac{s_2{}^2}{n_2}}}$$

# Hypothesis Testing

Problems with the above procedure

a) How do you calculate a confidence level?

b) What is an acceptable confidence level ?

- How would we test the hypothesis "the space shuttle is safe?"
- Is 1 explosion per 10 launches safe? Or 1 explosion per 100 launches?

A working definition of the confidence level:

*The probability of the event happening by chance.*

# Hypothesis Testing

Example: Suppose we measure some quantity (X) and we know that it is described by a gaussian *pdf* with μ=0 and σ=1. What is the confidence level for measuring X ≥2 (i.e. ≥2σ from the mean)?

$$P(X \geq 2) = \int_{2}^{\infty} P(\mu, \sigma, x)dx = \int_{2}^{\infty} P(0,1, x)dx = \frac{1}{\sqrt{2\pi}} \int_{2}^{\infty} e^{-\frac{x^2}{2}} dx = 0.023$$

Thus we would say that the confidence level for measuring X≥2 is 0.023 or 2.3% and we would expect to get a value of X ≥2 one out of about 40 tries if the underlying *pdf* is Gaussian.

# Hypothesis Testing

We wish to test if a quantity we have measured (m=average of n measurements ) is consistent with a known mean ($\mu_0$).

| Test | Conditions | Test Statistic | Test Distribution |
|---|---|---|---|
| $\mu = \mu_o$ | $\sigma^2$ known | $\frac{\mu - \mu_o}{\sigma / \sqrt{n}}$ | Gaussian<br>avg.= 0, $\sigma = 1$ |
| $\mu = \mu_o$ | $\sigma^2$ unknown | $\frac{\mu - \mu_o}{s / \sqrt{n}}$ | $t(n-1)$<br>student's "t-distribution" with n-1 DOF |

# Hypothesis Testing

*Example:* Do free quarks exist? Quarks are nature's fundamental building blocks and are thought to have electric charge (|q|) of either (1/3)e or (2/3)e (e = charge of electron). Suppose we do an experiment to look for |q| = 1/3 quarks.

We measure: q = 0.90 ± 0.2 This gives $\mu$ and $\sigma$

Quark theory: q = 0.33 This is $\mu_o$

We want to test the hypothesis $\mu = \mu_o$ when $\sigma$ is known. Thus we use the first line in the table.

$$z = \frac{\mu - \mu_o}{\sigma / \sqrt{n}} = \frac{0.9 - 0.33}{0.2 / \sqrt{1}} = 2.85$$

We want to calculate the probability for getting a $z \geq 2.85$, assuming a Gaussian *pdf*.

$$prob(z \geq 2.85) = \int_{2.85}^{\infty} P(\mu, \sigma, x)\,dx = \int_{2.85}^{\infty} P(0,1,x)dx = \frac{1}{\sqrt{2\pi}} \int_{2.85}^{\infty} e^{-\frac{x^2}{2}}\,dx = 0.002$$

The CL here is just 0.2 %! What we are saying here is that if we repeated our experiment 1000 times then the results of 2 of the experiments would measure a value q ≥ 0.9 if the true mean was q = 1/3. This is not strong evidence!

# Hypothesis Testing

Do charge 2/3 quarks exist?

$\boxed{H_0: 0.9 \pm 0.2 = 0.67}$

If instead of q = 1/3 quarks we tested for q = 2/3 what would we get for the CL?

Now we have $\mu = 0.9$ and $\sigma = 0.2$ as before but $\mu_O = 2/3$.

We now have $z = 1.17$ and prob($z \geq 1.17$) = 0.13 and the CL = 13%.

Now free quarks are starting to get believable!

If acceptable CL=5%, then we would accept $H_0$

# Hypothesis Testing

- Tests when both means are unknown but come from a gaussian *pdf*:

| Test | Conditions | Test Statistic | Test Distribution |
|---|---|---|---|
| $\mu_1 - \mu_2 = 0$ | $\sigma_1^2$ and $\sigma_2^2$ known | $\dfrac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}}$ | Gaussian<br>avg.= 0, $\sigma = 1$ |
| $\mu_1 - \mu_2 = 0$ | $\sigma_1^2 = \sigma_2^2 = \sigma^2$<br>unknown | $\dfrac{\mu_1 - \mu_2}{Q \cdot \sqrt{1 / n + 1 / m}}$<br>$Q^2 \equiv \dfrac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$ | $t\ (n+m-2)$ |
| $\mu_1 - \mu_2 = 0$ | $\sigma_1^2 \neq \sigma_2^2$<br>unknown | $\dfrac{\mu_1 - \mu_2}{\sqrt{s_1^2 / n + s_2^2 / m}}$ | approx. Gaussian<br>avg.= 0, $\sigma = 1$ |

n and m are the number of measurements for each mean

# Hypothesis Testing

## Example:

Do two experiments agree with each other?

Experiment A measures the $\Lambda_c$ lifetime to be (180 ± 7)fs while Experiment B measures (198 ± 7)fs.

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} = \frac{198 - 180}{\sqrt{(7)^2 + (7)^2}} = 1.82$$

$$\boxed{H_0\text{: } 180 \pm 7 = 198 \pm 7}$$

$$P(|z| \geq 1.82) = 1 - \int_{-1.82}^{1.82} P(\mu,\sigma,x)dx = 1 - \int_{-1.82}^{1.82} P(0,1,x)dx = 1 - \frac{1}{\sqrt{2\pi}} \int_{-1.82}^{1.82} e^{-\frac{x^2}{2}} dx = 1 - 0.93 = 0.07$$

Thus 7% of the time we should expect the experiments to disagree at this level.

51

If acceptable CL=5%, then we would accept $H_0$

# Hypothesis Testing

- Example: We compare results of two independent experiments to see if they agree with each other.
  - ○ Exp. 1 1.00 ± 0.01
  - ○ Exp. 2 1.04 ± 0.02

Use the first line of the table and set $n = m = 1$.

$$z = \frac{x_1 - x_2}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} = \frac{1.04 - 1.00}{\sqrt{(0.01)^2 + (0.02)^2}} = 1.79$$

- $z$ is distributed according to a Gaussian with $\mu = 0$, $\sigma = 1$.
- Probability for the two experiments to disagree by $\geq |0.04|$:

$$prob(|z| \geq 1.79) = 1 - \int_{-1.79}^{1.79} P(\mu, \sigma, x)dx = 1 - \int_{-1.79}^{1.79} P(0,1,x)dx = 1 - \frac{1}{\sqrt{2\pi}} \int_{-1.79}^{1.79} e^{-\frac{x^2}{2}} dx = 0.07$$

We don't care which experiment has the larger result so we use ± $z$.

Thus 7% of the time we should expect the experiments to disagree at this level.

If acceptable CL=5%, then we would accept $H_0$

# Example: Weight Loss for Diet vs Exercise

*Did dieters lose more fat than the exercisers?*

Diet Only:

sample mean = 5.9 kg
sample standard deviation = 4.1 kg sample size, $n = 42$
standard error = SEM1 = $4.1/ \sqrt{42} = 0.633$

Exercise Only:
sample mean = 4.1 kg
sample standard deviation = 3.7 kg sample size = $n = 47$
standard error = SEM2 = $3.7/ \sqrt{47} = 0.540$
measure of variability = $[(0.633)2 + (0.540)2] = 0.83$

# Example: Weight Loss for Diet vs Exercise

- **Step 1. Determine the null and alternative hypotheses.**
- ***Null hypothesis:*** No difference in average fat lost in population for two methods. Population mean difference is ***zero***.
- ***Alternative hypothesis:*** There is a difference in average fat lost in population for two methods. Population mean difference is not ***zero***.

# Example: Weight Loss for Diet vs Exercise

- **Step 2. Collect and summarize data into a test statistic.**
- The sample mean difference = 5.9 – 4.1 = 1.8 kg and the standard error of the difference is 0.83.
- So the *test statistic: z* = 1.8 – 0/0.83 = 2.17

# Example: Weight Loss for Diet vs Exercise

- **Step 3. Determine the *p*-value.**
- ***Recall the alternative hypothesis was two-sided.***
- *p*-value = 2 × [proportion of bell-shaped curve above 2.17]
- Table => proportion is about 2 × 0.015 = 0.03.

# Example: Weight Loss for Diet vs Exercise

- **Step 4. Make a decision.**

The ***p*-value of 0.03 is less than or equal to 0.05**, so ...

- If really no difference between dieting and exercise as fat loss methods, would see such an extreme result only 3% of the time, or 3 times out of 100.
- Prefer to believe truth does not lie with null hypothesis. We conclude that there is a ***statistically significant difference between average fat loss for the two methods***.

# Public Opinion About President

- A newspaper reported the results of a public opinion poll that asked: *“From everything you know about xxx, does he have the honesty and integrity you expect in a president?”*
- Poll **surveyed 518 adults and 233, or 0.45** of them (clearly less than half), answered yes.
- Could presedent’s adversaries conclude from this that **only a minority (less than half) of the population** of Turkish thought xxx had the honesty and integrity to be president?

# Public Opinion About President

**Step 1. Determine the null and alternative hypotheses.**

- ***Null hypothesis:*** There is no clear winning opinion on this issue; the proportions who would answer yes or no are each 0.50.
- ***Alternative hypothesis:*** Fewer than 0.50, or 50%, of the population would answer yes to this question. The majority do not think xxx has the honesty and integrity to be president.

# Public Opinion About President

**Step 2. Collect and summarize data into a test statistic.**

- Sample proportion is: 233/518 = 0.45.
- The *standard deviation* = (0.50) × (1 – 0.50)/518 = 0.022.
- *Test statistic: z* = (0.45 – 0.50)/0.022 = –2.27

# Public Opinion About President

## Step 3. Determine the *p*-value.

- ***Recall the alternative hypothesis was one-sided.***
- *p*-value = proportion of bell-shaped curve below
- –2.27 Exact *p*-value = 0.0116.

## Step 4. Make a decision.

- The ***p*-value of 0.0116 is less than 0.05**, so we conclude that the proportion of Turkish adults who believed xxx had the honesty and integrity they expected in a president was **significantly less** than a majority.

# Quitting Smoking with Nicotine Patches

- Compared the smoking cessation rates for smokers randomly assigned to use a nicotine patch versus a placebo patch.
- ***Null hypothesis:*** The proportion of smokers in the population who would quit smoking using a nicotine patch and a placebo patch are the **same**.
- ***Alternative hypothesis:*** The proportion of smokers in the population who would quit smoking using a **nicotine patch is higher** than the proportion who would quit using a placebo patch.

# Homework or not

- Step 1. Determine the null and alternative hypotheses.
- Null hypothesis: The proportion of students at the university who oppose homework assignment is 0.50.
- Alternative hypothesis: The proportion of students at university who oppose homework assignment is greater than 0.50.
- Step 2. Collect data and summarize with test statistic. In a random sample of 400 students, 220 or 55% oppose.
- So the standard deviation = (0.50) × (1 – 0.50)/400 = 0.025.
- Test statistic: z = 0.55 – 0.50/0.025 = 2.00

# Homework or not

## Step 3. Determine the *p*-value.

- ***Recall the alternative hypothesis:*** The proportion of students at university who oppose homework assignment  is **greater than** 0.50.
- So *p*-value = proportion of bell-shaped curve *above* 2.00. Table 8.1 => proportion is between 0.02 and 0.025. Using computer/calculator: exact *p*-value = 0.0228.

## Step 4. Make a decision.

The ***p*-value is less than or equal to 0.05**, so we conclude:

- Reject the null hypothesis
- Accept the alternative hypothesis
- The true population proportion opposing homework assignment is significantly *greater* than 0.50.

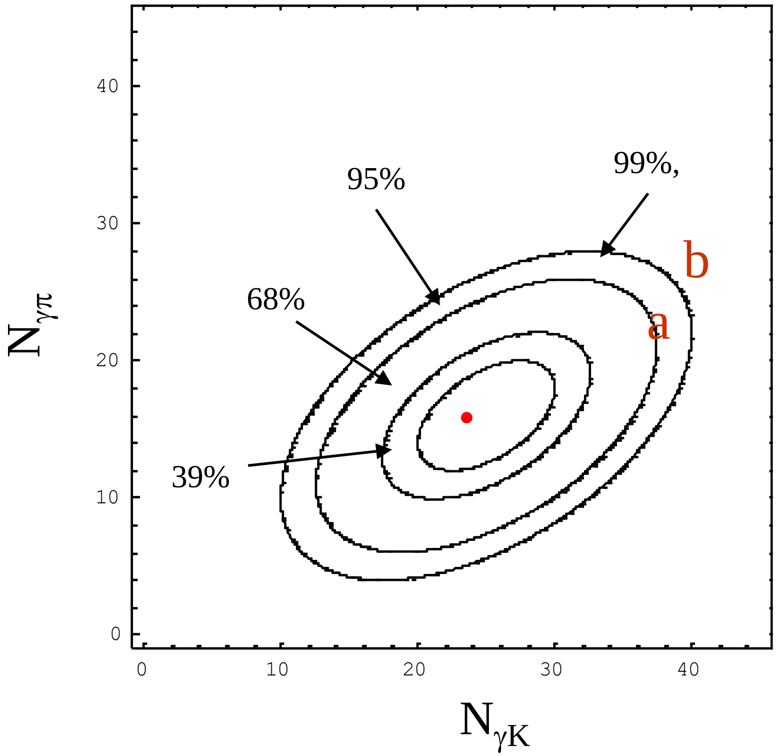
# In Class Exercise

Suppose in an experiment a maximum likelihood analysis to search for B→$\gamma\pi^-$ and B→$\gamma K^-$ events. The results of the MLM fit are:

$N_{\gamma\pi}$ =16±4, $N_{\gamma K}$ =25±5, $\rho$=0.5

$N_{\gamma\pi}$ and $N_{\gamma K}$ are highly correlated, since at high momentum (>2GeV) it has a hard to separate π's and K's.

Writting a python script obtain exclusion contours below



# Example: Python

```
import plotly.plotly as py
import plotly.graph_objs as go
from plotly.tools import FigureFactory as FF
import numpy as np
import pandas as pd
import scipy

data1 = np.random.normal(0, 1, size=50)
data2 = np.random.normal(2, 1, size=50)
```

# Example: Python

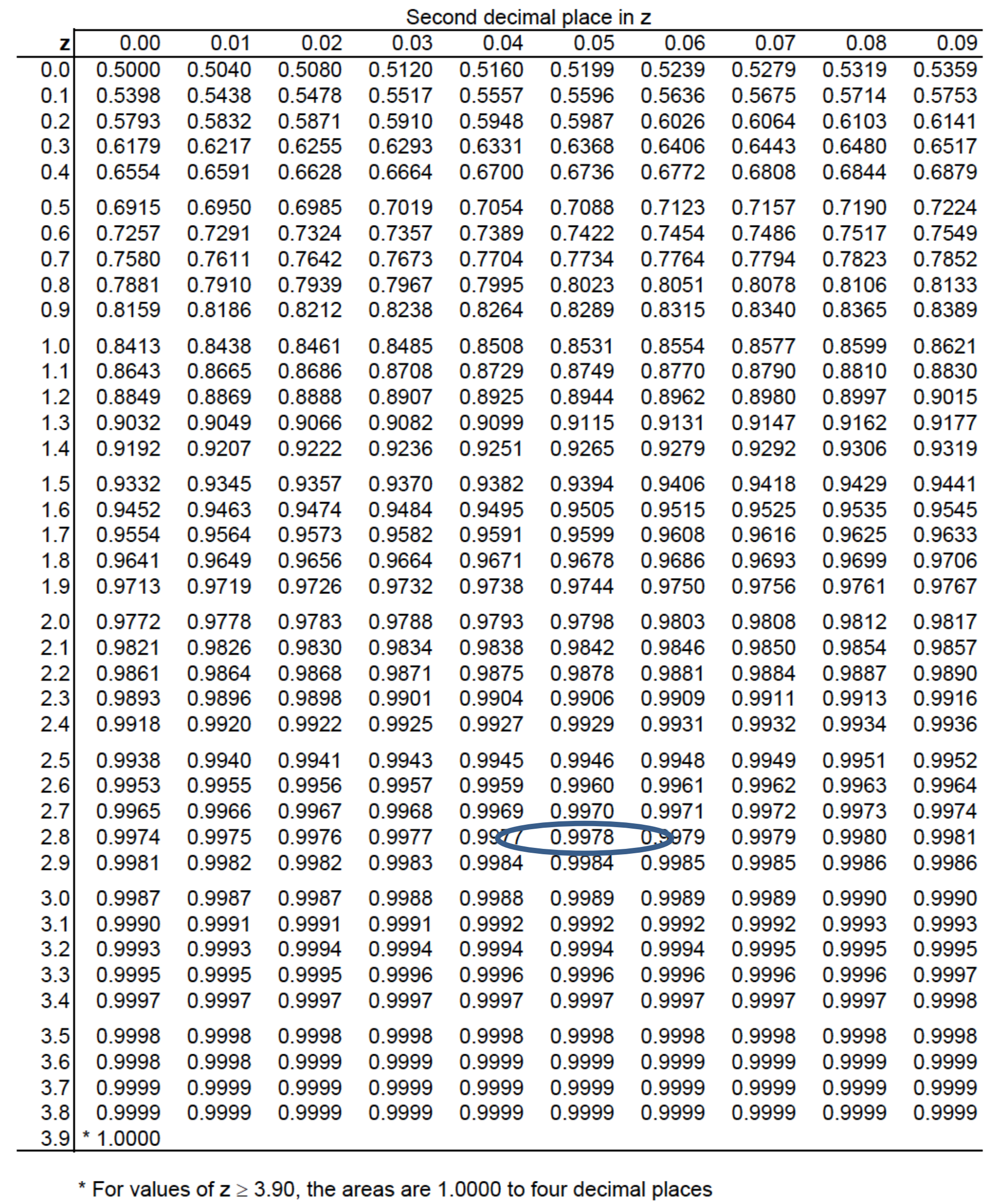
```
x = np.linspace(-4, 4, 160)
y1 = scipy.stats.norm.pdf(x)
y2 = scipy.stats.norm.pdf(x, loc=2)
trace1 = go.Scatter(
    x = x,
    y = y1,
    mode = 'lines+markers',
    name='Mean of 0'
)
trace2 = go.Scatter(
    x = x,
    y = y2,
    mode = 'lines+markers',
    name='Mean of 2'
)
data = [trace1, trace2]
py.iplot(data, filename='normal-dists-plot')
```

# Example: Python

```
true_mu = 0

onesample_results = scipy.stats.ttest_1samp(data1, true_mu)

matrix_onesample = [
    ['', 'Test Statistic', 'p-value'],
    ['Sample Data', onesample_results[0], onesample_results[1]]
]

onesample_table = FF.create_table(matrix_onesample, index=True)
py.iplot(onesample_table, filename='onesample-table')
```

Since our p-value is greater than our Test-Statistic, we have good evidence to not reject the null-hypothesis at the 0.05 significance level.

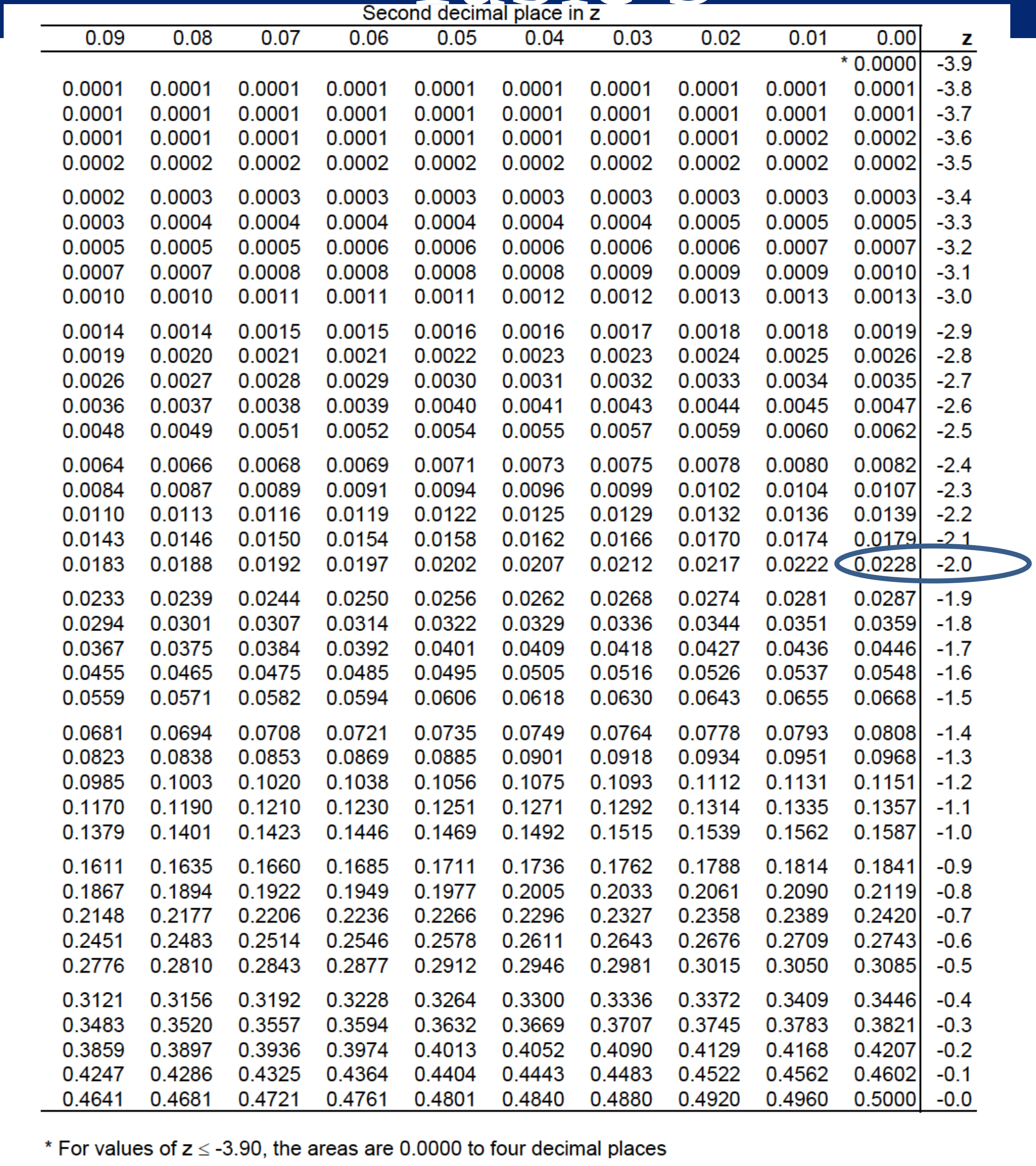
# Table 3



|  | Second decimal place in z |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |
| 3.1 | 0.9990 | 0.9991 | 0.9991 | 0.9991 | 0.9992 | 0.9992 | 0.9992 | 0.9992 | 0.9993 | 0.9993 |
| 3.2 | 0.9993 | 0.9993 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9995 | 0.9995 | 0.9995 |
| 3.3 | 0.9995 | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9997 |
| 3.4 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |
| 3.5 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 |
| 3.6 | 0.9998 | 0.9998 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.7 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.8 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.9 | * 1.0000 |  |  |  |  |  |  |  |  |  |

* For values of z ≥ 3.90, the areas are 1.0000 to four decimal places

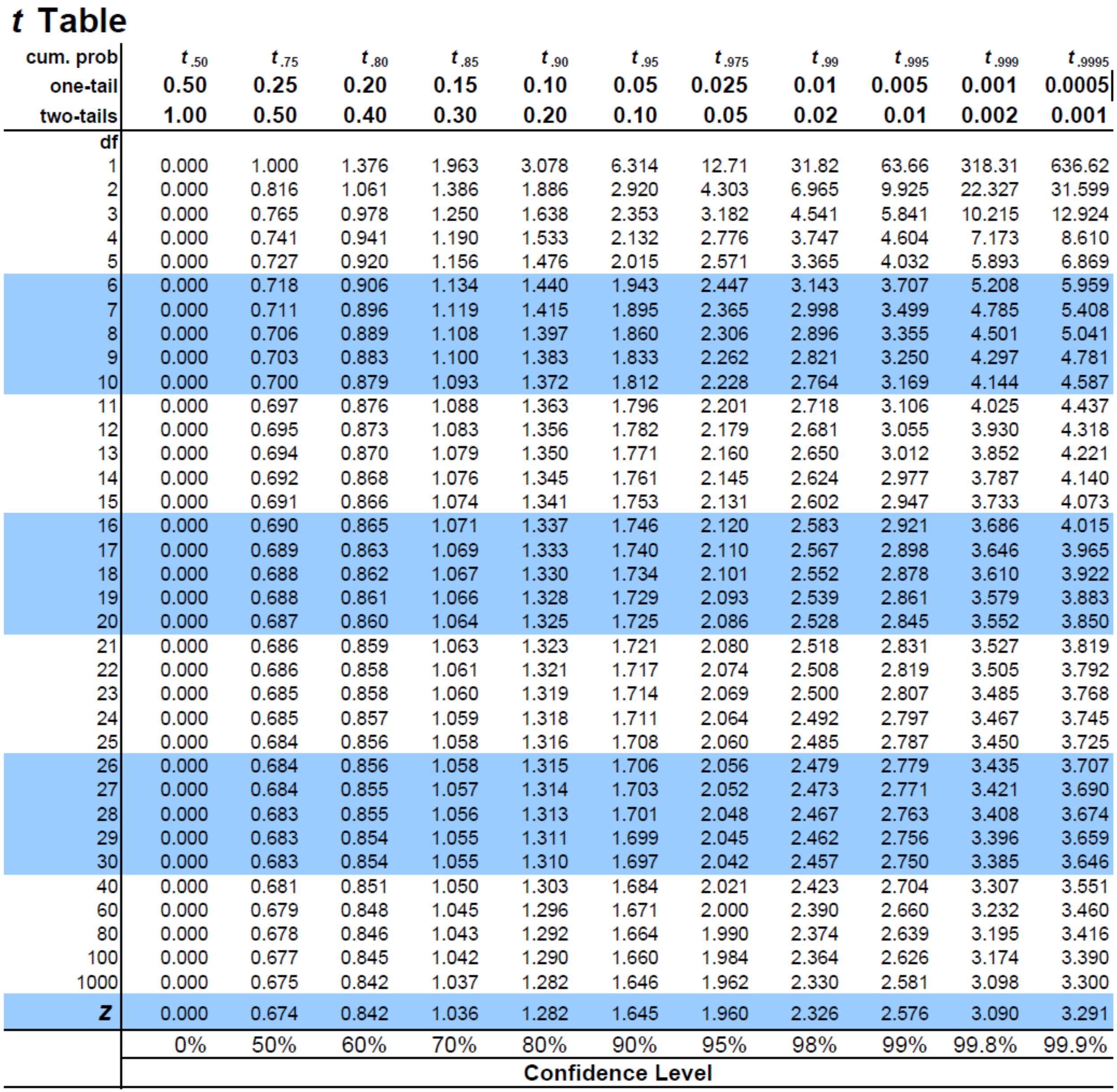
# Table 3



| Second decimal place in z | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.09 | 0.08 | 0.07 | 0.06 | 0.05 | 0.04 | 0.03 | 0.02 | 0.01 | 0.00 | z |
| | | | | | | | | | * 0.0000 | -3.9 |
| 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | -3.8 |
| 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | -3.7 |
| 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 | -3.6 |
| 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | -3.5 |
| 0.0002 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | -3.4 |
| 0.0003 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0005 | 0.0005 | 0.0005 | -3.3 |
| 0.0005 | 0.0005 | 0.0005 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0007 | 0.0007 | -3.2 |
| 0.0007 | 0.0007 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0010 | -3.1 |
| 0.0010 | 0.0010 | 0.0011 | 0.0011 | 0.0011 | 0.0012 | 0.0012 | 0.0013 | 0.0013 | 0.0013 | -3.0 |
| 0.0014 | 0.0014 | 0.0015 | 0.0015 | 0.0016 | 0.0016 | 0.0017 | 0.0018 | 0.0018 | 0.0019 | -2.9 |
| 0.0019 | 0.0020 | 0.0021 | 0.0021 | 0.0022 | 0.0023 | 0.0023 | 0.0024 | 0.0025 | 0.0026 | -2.8 |
| 0.0026 | 0.0027 | 0.0028 | 0.0029 | 0.0030 | 0.0031 | 0.0032 | 0.0033 | 0.0034 | 0.0035 | -2.7 |
| 0.0036 | 0.0037 | 0.0038 | 0.0039 | 0.0040 | 0.0041 | 0.0043 | 0.0044 | 0.0045 | 0.0047 | -2.6 |
| 0.0048 | 0.0049 | 0.0051 | 0.0052 | 0.0054 | 0.0055 | 0.0057 | 0.0059 | 0.0060 | 0.0062 | -2.5 |
| 0.0064 | 0.0066 | 0.0068 | 0.0069 | 0.0071 | 0.0073 | 0.0075 | 0.0078 | 0.0080 | 0.0082 | -2.4 |
| 0.0084 | 0.0087 | 0.0089 | 0.0091 | 0.0094 | 0.0096 | 0.0099 | 0.0102 | 0.0104 | 0.0107 | -2.3 |
| 0.0110 | 0.0113 | 0.0116 | 0.0119 | 0.0122 | 0.0125 | 0.0129 | 0.0132 | 0.0136 | 0.0139 | -2.2 |
| 0.0143 | 0.0146 | 0.0150 | 0.0154 | 0.0158 | 0.0162 | 0.0166 | 0.0170 | 0.0174 | 0.0179 | -2.1 |
| 0.0183 | 0.0188 | 0.0192 | 0.0197 | 0.0202 | 0.0207 | 0.0212 | 0.0217 | 0.0222 | 0.0228 | -2.0 |
| 0.0233 | 0.0239 | 0.0244 | 0.0250 | 0.0256 | 0.0262 | 0.0268 | 0.0274 | 0.0281 | 0.0287 | -1.9 |
| 0.0294 | 0.0301 | 0.0307 | 0.0314 | 0.0322 | 0.0329 | 0.0336 | 0.0344 | 0.0351 | 0.0359 | -1.8 |
| 0.0367 | 0.0375 | 0.0384 | 0.0392 | 0.0401 | 0.0409 | 0.0418 | 0.0427 | 0.0436 | 0.0446 | -1.7 |
| 0.0455 | 0.0465 | 0.0475 | 0.0485 | 0.0495 | 0.0505 | 0.0516 | 0.0526 | 0.0537 | 0.0548 | -1.6 |
| 0.0559 | 0.0571 | 0.0582 | 0.0594 | 0.0606 | 0.0618 | 0.0630 | 0.0643 | 0.0655 | 0.0668 | -1.5 |
| 0.0681 | 0.0694 | 0.0708 | 0.0721 | 0.0735 | 0.0749 | 0.0764 | 0.0778 | 0.0793 | 0.0808 | -1.4 |
| 0.0823 | 0.0838 | 0.0853 | 0.0869 | 0.0885 | 0.0901 | 0.0918 | 0.0934 | 0.0951 | 0.0968 | -1.3 |
| 0.0985 | 0.1003 | 0.1020 | 0.1038 | 0.1056 | 0.1075 | 0.1093 | 0.1112 | 0.1131 | 0.1151 | -1.2 |
| 0.1170 | 0.1190 | 0.1210 | 0.1230 | 0.1251 | 0.1271 | 0.1292 | 0.1314 | 0.1335 | 0.1357 | -1.1 |
| 0.1379 | 0.1401 | 0.1423 | 0.1446 | 0.1469 | 0.1492 | 0.1515 | 0.1539 | 0.1562 | 0.1587 | -1.0 |
| 0.1611 | 0.1635 | 0.1660 | 0.1685 | 0.1711 | 0.1736 | 0.1762 | 0.1788 | 0.1814 | 0.1841 | -0.9 |
| 0.1867 | 0.1894 | 0.1922 | 0.1949 | 0.1977 | 0.2005 | 0.2033 | 0.2061 | 0.2090 | 0.2119 | -0.8 |
| 0.2148 | 0.2177 | 0.2206 | 0.2236 | 0.2266 | 0.2296 | 0.2327 | 0.2358 | 0.2389 | 0.2420 | -0.7 |
| 0.2451 | 0.2483 | 0.2514 | 0.2546 | 0.2578 | 0.2611 | 0.2643 | 0.2676 | 0.2709 | 0.2743 | -0.6 |
| 0.2776 | 0.2810 | 0.2843 | 0.2877 | 0.2912 | 0.2946 | 0.2981 | 0.3015 | 0.3050 | 0.3085 | -0.5 |
| 0.3121 | 0.3156 | 0.3192 | 0.3228 | 0.3264 | 0.3300 | 0.3336 | 0.3372 | 0.3409 | 0.3446 | -0.4 |
| 0.3483 | 0.3520 | 0.3557 | 0.3594 | 0.3632 | 0.3669 | 0.3707 | 0.3745 | 0.3783 | 0.3821 | -0.3 |
| 0.3859 | 0.3897 | 0.3936 | 0.3974 | 0.4013 | 0.4052 | 0.4090 | 0.4129 | 0.4168 | 0.4207 | -0.2 |
| 0.4247 | 0.4286 | 0.4325 | 0.4364 | 0.4404 | 0.4443 | 0.4483 | 0.4522 | 0.4562 | 0.4602 | -0.1 |
| 0.4641 | 0.4681 | 0.4721 | 0.4761 | 0.4801 | 0.4840 | 0.4880 | 0.4920 | 0.4960 | 0.5000 | -0.0 |

* For values of z ≤ -3.90, the areas are 0.0000 to four decimal places

# Table 4



## *t* Table

| cum. prob | $t_{.50}$ | $t_{.75}$ | $t_{.80}$ | $t_{.85}$ | $t_{.90}$ | $t_{.95}$ | $t_{.975}$ | $t_{.99}$ | $t_{.995}$ | $t_{.999}$ | $t_{.9995}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **one-tail** | **0.50** | **0.25** | **0.20** | **0.15** | **0.10** | **0.05** | **0.025** | **0.01** | **0.005** | **0.001** | **0.0005** |
| **two-tails** | **1.00** | **0.50** | **0.40** | **0.30** | **0.20** | **0.10** | **0.05** | **0.02** | **0.01** | **0.002** | **0.001** |
| **df** | | | | | | | | | | | |
| 1 | 0.000 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.31 | 636.62 |
| 2 | 0.000 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.000 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.000 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.000 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.000 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | 0.000 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.408 |
| 8 | 0.000 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | 0.000 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | 0.000 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | 0.000 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | 0.000 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | 0.000 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | 0.000 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | 0.000 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | 0.000 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | 0.000 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | 0.000 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| 19 | 0.000 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | 0.000 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | 0.000 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | 0.000 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | 0.000 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| 24 | 0.000 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | 0.000 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | 0.000 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | 0.000 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | 0.000 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | 0.000 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | 0.000 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |
| 40 | 0.000 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 3.551 |
| 60 | 0.000 | 0.679 | 0.848 | 1.045 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 3.460 |
| 80 | 0.000 | 0.678 | 0.846 | 1.043 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 3.416 |
| 100 | 0.000 | 0.677 | 0.845 | 1.042 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 3.390 |
| 1000 | 0.000 | 0.675 | 0.842 | 1.037 | 1.282 | 1.646 | 1.962 | 2.330 | 2.581 | 3.098 | 3.300 |
| ***z*** | 0.000 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | 3.291 |
| | 0% | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99.8% | 99.9% |
| | | | | | **Confidence Level** | | | | | | |

# Probabilities Associated with Errors

**The Power of a Test**

The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true. If the population value falls close to the value specified in null hypothesis, then it is difficult to get enough evidence from the sample to conclusively choose the alternative hypothesis.

**When to Reject the Null Hypothesis**

In deciding whether to reject the null hypothesis consider the consequences of the two potential types of errors.

- If consequences of a type 1 error are very serious, then only reject null hypothesis if the *p*-value is very small.
- If type 2 error more serious, should be willing to reject null hypothesis with a moderately large *p*-value, 0.05 to 0.10.

# Significance & Power

How often a type I error occurs is called the significance of the test. Often called α;
α should be small: 1% or 5%
– Often discrete result, thus there is no region for exactly 5%, but for example 4.7%
– Composite hypothesis have to be evaluated for the "worst" point in parameter-space.

How often a type II error occurs is called β. The power of a test is defined as (1-β).
Ideal Test: Both α and β are small => small significance and large power.

# Hypothesis Testing

A procedure for using hypothesis testing:

a) Measure (or calculate) parameter

b) Take a theory that you wish to compare with your measurement (theory, experiment)

c) Form a hypothesis (e.g. my measurement, x, is consistent with the PDG value) $H_0$: $x=x_{PDG}$

$H_0$ is called the "null hypothesis"

d) Calculate the confidence level that the hypothesis is true

e) Accept or reject the hypothesis depending on some minimum acceptable confidence level

# P-value

- The P-value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to $H_0$ as the value calculated from the available sample.

- One advantage is that the P-value provides an intuitive measure of the strength of evidence in the data against $H_0$.

- The P-value is not the probability that $H_0$ is true, nor is it an error probability!

- To determine the P-value, we must first decide which values of the test statistic are at least as contradictory to $H_0$ as the value obtained from our sample.

# Probabilities Associated with Errors

- We can only specify the conditional probability of making a type 1 error, given that the null hypothesis is true. That probability is called the **level of significance**, usually 0.05.

**Level of Significance and Type I Errors**

*If null hypothesis is true, probability of making a type 1 error is equal to the level of significance, usually 0.05.*

*If null hypothesis is not true, a type 1 error cannot be made.*

**Type 2 Errors**

*A type 2 error is made if the alternative hypothesis is true, but you fail to choose it. The probability of doing that depends on which part of the alternative hypothesis is true, so computing the probability of making a type 2 error is not feasible.*

# Hypothesis Testing

You must know the underlying *pdf* to calculate the limits.

Example: suppose we have a scale of known accuracy ($\sigma$ = 10 gm ) and we weigh something to be 20 gm. Assuming a gaussian *pdf* we could calculate a 2.3% chance that our object weighs ≤ 0 gm?? We must make sure that the probability distribution is defined in the region where we are trying to extract information.

# Example: Hypothesis Testing

1. Determine the **null** hypothesis and the **alternative** hypothesis.

2. Collect and summarize the **data** into a **test statistic**.

3. Use the test statistic to determine the ***p*-value**.

4. The result is **statistically significant** if the *p*-value is less than or equal to the level of significance.

# Example: Hypothesis Testing

- If the null and alternative hypotheses are expressed in terms of a **population proportion, mean, or difference between two means** and if the sample sizes are large ...

- ... the **test statistic** is simply the corresponding **standardized score** computed assuming the null hypothesis is true; and the ***p*-value** is found from a table of percentiles for standardized scores.