# EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

*ed.kroc@ubc.ca*

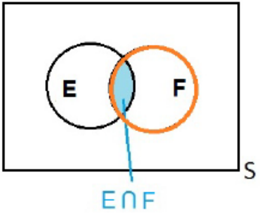
January 16, 2020

# Last time

- Intro to experimental design
- Basics of probability
- Conditional probability

# Conditional Probability

- In general, given that an event $F$ has occurred, the probability that another event $E$ occurs is called the *conditional probability of $E$ given $F$*.




- Notation and formula:

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E \text{ and } F)}{\Pr(F)}$$

# Independence of Events

**Definition**

Two events $E$ and $F$ are said to be **independent** if and only if $\Pr(E \mid F) = \Pr(E)$ or $\Pr(F \mid E) = \Pr(F)$.

- By definition of conditional probability then, we have

$$\Pr(E \cap F) = \Pr(E) \cdot \Pr(F) \quad \text{if and only if } E,\ F \text{ are independent.}$$

- This definition matches with our intuition: if two events are independent, then the fact that one event happens should *not* have any affect on how likely the other event is to happen.

# The Prosecutor's Fallacy

The Prosecutor's Fallacy is a common probability *misconception*: the fallacy is thinking that $\Pr(A \cap B)$ is the same as $\Pr(A \mid B)$.

- This is obviously false! Only true if $\Pr(B) = 1$ or if $\Pr(A \cap B) = 0$. Recall that
$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

- This fallacy is quite common and can have many distressing consequences...

# The Case of Sally Clark

- In 1998, Sally Clark was accused of murdering her two infant sons. One died in 1996 at eleven weeks old. The second died a year later at eight weeks of age.
- Sir Roy Meadow, pediatrician and expert witness for the prosecution, testified that the chance of two children in the same family dying from Sudden Infant Death Syndrome (SIDS) was about $(1/8500)^2$, or 1 in 73 million.
- On the strength of this testimony alone, Clark was convicted in 1999. The Royal Statistical Society then pointed out the flaws in the argument. What are they?

# The Case of Sally Clark

- Flaw #1: The events of two *siblings* dying from SIDS are *not* independent. There is a genetic component! In reality, the probability of two children from the same family dying of SIDS is much closer to $1/8500$ than to $(1/8500)^2$.

- Flaw #2: Meadow confused the conditional and unconditional probabilities (the Prosecutor's Fallacy).

  Let $I$ : event that Clark is innocent of murder, $E$ : event of two dead children (the evidence).

  We know that in general,

$$\Pr(I \mid E) \neq \Pr(E \text{ and } I).$$

# The Case of Sally Clark

Now,

$$\begin{aligned}\Pr(I \mid E) &= \frac{\Pr(I \text{ and } E)}{\Pr(E)} \\ &= \frac{\Pr(I \text{ and } E)}{\Pr(\{I \text{ and } E\} \text{ or } \{I^c \text{ and } E\})} \\ &= \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)}\end{aligned}$$

What are the events $I$ and $E$ and $I^c$ and $E$?

- $I$ and $E$ is the event of the two chidren dying by SIDS.
- $I^c$ and $E$ is the event of the two children dying by murder.

Double SIDS is rare, but double murder is much, much rarer! So,

$$\Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E).$$

$$\Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E)$$

means:

$$\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E) + \Pr(I \text{ and } E)$$

$$\frac{1}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)} \gg \frac{1}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)}$$

$$\frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)} \gg \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)}$$

$$\Pr(I \mid E) \gg \frac{1}{2}$$

# The Case of Sally Clark

$$\Pr(I^c \text{ and } E) \ll \Pr(I \text{ and } E)$$

means:

$$\Pr(I \mid E) = \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I^c \text{ and } E)}$$
$$\gg \frac{\Pr(I \text{ and } E)}{\Pr(I \text{ and } E) + \Pr(I \text{ and } E)} = \frac{1}{2}$$

So, $\Pr(I \mid E) \approx 1$!

Moral of the story 1: **circumstantial evidence of a rare event is very weak evidence.**

Moral of the story 2: **conditional information is radically different from unconditional information.**

# Today

- Random variables
- Means (expectations), variance, standard deviation
- Bernoulli, Binomial, and Normal random variables
- Sample statistics (standard errors, confidence intervals, CLT)
- Hypothesis testing and p-values

# Random Variables

## Definition

A **random variable** is a function that maps events in a sample space to the real numbers. We use uppercase letters to denote a random variable, and lowercase letters to denote sample realizations of that random variable.

# Random Variables

## Definition

A **random variable** is a function that maps events in a sample space to the real numbers. We use uppercase letters to denote a random variable, and lowercase letters to denote sample realizations of that random variable.

Example: Toss a fair coin three times:

$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

We can define a random variable $X$ to be the number of heads observed in the three tosses:

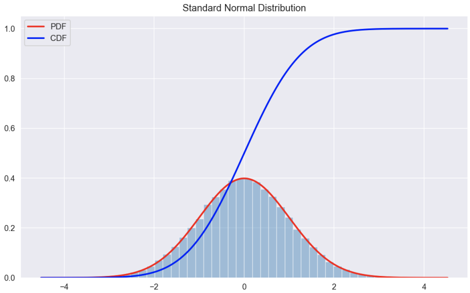
| Event | HHH | HHT | HTH | THH | HTT | THT | TTH | TTT |
|---|---|---|---|---|---|---|---|---|
| $X = x$ | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 0 |

# Random Variables

- All r.v.'s come equipped with a *cumulative distribution function (CDF)* that lets us figure out probabilties of events.
- The CDF simply *accumulates* probabilties up to a certain value:

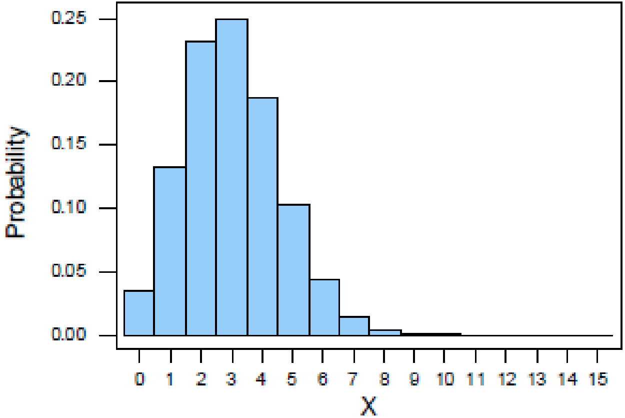
$$\Pr(X \leqslant x),$$

where $X$ is the random variable, and $x \in \mathbb{R}$.



- Discrete r.v.'s also have a *probability mass function (PMF)* that tells us the probabilities of single events:

$$\Pr(X = x).$$

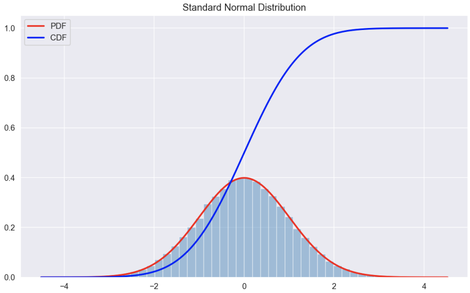


# Random Variables

- Continuous r.v.'s instead have a *probability density function (PDF)*, denoted $f(x)$, that allows us to write:

$$\Pr(a \leqslant X \leqslant b) = \int_a^b f(x)dx,$$

for any $a, b \in \mathbb{R}$.



# Expectation of a Random Variable

## Definition

The **expectation** of $X$ (discrete) is given by

$$\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x).$$

The **expectation** of $X$ (continuous) is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx.$$

The expectation of $X$ is also referred to as the **expected value** of $X$ or the **mean** of $X$. We often denote the mean by $\mu$ or $\mu_X$.

Note: the *expectation* generalizes the idea of the *simple average* of a bunch of numbers.

# Variance and Standard Deviation of a Random Variable

## Definition

We define the **variance** of a r.v. $X$ as

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2].$$

For $X$ discrete, this is

$$\text{Var}(X) = \sum_x (x - \mu)^2 \Pr(X = x).$$

For $X$ continuous, this is

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

Note: the *variance* of a random variable quantifies how likely it is that the random variable takes on values *away* from its mean/expectation.

# Variance and Standard Deviation of a Random Variable

## Definition

The **standard deviation** of $X$ is

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

The standard deviation has the *same units* as the random variable itself.

- We often denote the variance of $X$ by $\sigma^2$ or $\sigma_X^2$, and the standard deviation of $X$ by $\sigma$ or $\sigma_X$.
- Standard deviations give the same information as variances (just different units).
- Standard deviations are easier to interpret, but variances are easier to work with mathematically.

# Bernoulli trials

- A Bernoulli trial is a random experiment that gives only one of two outcomes, usually referred to as "success" and "failure".

  Examples:
  - Toss a fair coin once: "success" if a head is tossed, "failure" if a tail is tossed.
  - Medical diagnostic: "success" if patient has disease, "faliure" if patient is healthy.

- The number of "successes" in a Bernoulli trial (either 0 or 1) is a Bernoulli random variable with parameter $p$, where $p$ is the probability of the "success" outcome, $0 \leqslant p \leqslant 1$. We use the notation $X \sim Bernoulli(p)$ or $X \sim Ber(p)$ to denote that $X$ is distributed according to a Bernoulli random variable with parameter $p$.

# Binomial Random Variables

- A Binomial experiment consists of $n$ (fixed in advance) *identical and independent* Bernoulli trials.

  Examples:
  - Toss a fair coin $n = 10$ times: "success" if a head is tossed, "failure" if a tail is tossed.
  - Medical diagnostics: run the same test on $n = 100$ patients with equal chance of disease: "success" if patient has disease, "faliure" if patient is healthy.

- Let the $n$ Bernoulli trials be given by $X_1, X_2, \ldots, X_n$, where $X_i \sim Ber(p)$.

- Define $Y = X_1 + X_2 + \cdots + X_n$.
  $Y$ is the total number of successes out of the $n$ trials.
  $Y$ is a Binomial random variable with parameters $n$ and $p$, denoted $Y \sim Bin(n, p)$.

# Normal Random Variables

- Recall that a probability density function (PDF) for a continuous random variable is a function that tells us how to calculate the likelihood of different outcomes for the random variable.

- A random variable $X$ that follows the Normal, or Gaussian, distribution has a PDF given by

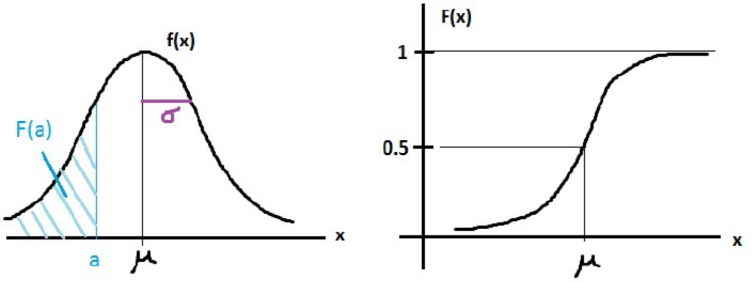
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

  Definition of PDF implies: $\Pr(a \leqslant X \leqslant b) = \int_a^b f(x)dx$.

- We write $X \sim N(\mu, \sigma^2)$, where $\mu$ is the mean parameter, and $\sigma^2$ is the variance parameter.

# Normal Random Variables

The Normal density is the classic "bell curve", and is perfectly symmetrical about the mean $\mu$.




- A standard Normal random variable $Z$ is a Normal random variable with $\mu = 0$ and $\sigma^2 = 1$: $Z \sim N(0, 1)$.

# Standardizing Random Variables

## Proposition

*Let $X \sim N(\mu, \sigma^2)$. Then the random variable $Z = \frac{X-\mu}{\sigma}$ is a standard Normal random variable; i.e. $Z \sim N(0, 1)$.*

- In general, the process of transforming a random variable by subtracting its mean and then dividing by its standard deviation is called *standardizing*.
- The resulting transformed random variable is called a *standardized* random variable. It will *always* have mean 0 and standard deviation 1.
- This allows us to make meaningful comparisons that do *not* depend on units of measurement.

# Applications of Normal Random Variables

- The Central Limit Theorem forms the backbone of the theory behind many classical analytical tools in statistics:
  - Tests of hypotheses about sample means (e.g. z-tests and t-tests)
  - Analysis of variance (ANOVA)
  - Linear regression
- The Normal distribution is often applied to analyze errors in measurement (e.g. random errors in making astronomical observations)
- The Normal distribution is often a great approximation to real world variables, e.g. height, weight, body temperature.
- The Normal distribution is used to define a bunch of other random variables with further statistical and real world applications, e.g.:
  - Student's t-distribution
  - Chi-squared distribution
  - Fisher F-distribution

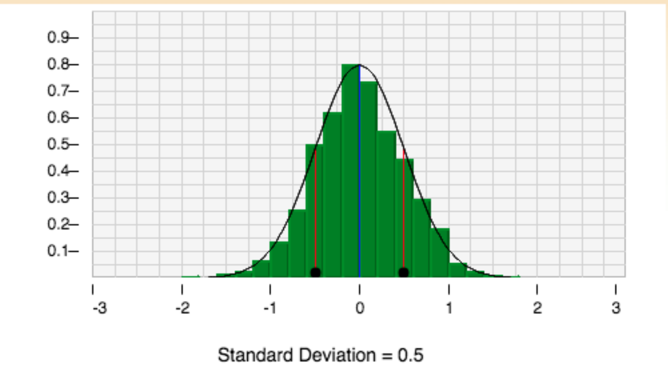
# Sample Statistics

- In practice, we study a random variable by observing its values on only a *sample*.
- Studying this sample allows us to infer properties of the actual random variable if the sample is random and representative.

- This is basically what applied statistics is all about!

# Sample Statistics

- We can approximate a r.v.'s PMF or PDF by plotting a *histogram* of our sample data.




Visit: http://www.shodor.org/interactivate/activities/NormalDistribution/

# Sample Statistics

- We can get a sense of the "typical" value of our r.v. by calculating a *sample mean*, *sample median*, or *sample mode*.

- Let $\{X_1, \ldots, X_n\}$ denote a random sample of $n$ independent observations from the random variable $X$. We define the **sample mean** by:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- sample median = 50th percentile of sample data

- sample mode = most commonly observed value in sample data

- Rememeber: these can all be different!

- We can get a sense of the spread or dispersion (variability) of our r.v. by calculating a *sample variance*.

- Let $\{X_1, \ldots, X_n\}$ denote a random sample of $n$ independent observations from the random variable $X$. We define the **sample variance** by:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

# Sample Statistics vs. Properties of Random Variables

- Although the definitions of *expectation* and *sample mean*, and of *variance* and *sample variance*, look very similar, they are fundamentally different.
    - Sample mean and variance are *functions of the data/sample*. Different samples will generate different values for sample mean/variance *even if the samples are from the same population*.
    - Expectactions and variances of random variables are idealized quantities. They are inherent properties of the random phenomenon we are studying. We usually cannot calculate them in practice; we can only estimate them via our *sample* approximations.

# Standard Errors

- Because sample statistics are random (i.e. not fixed) quantities, they are genuine random variables on their own!
- Thus, they have expectations, variances, std. devs. of their own.
- Terminology: the **standard error** of a sample statistic is simply its standard deviation.
- If $T$ denotes a sample statistic, then we usually write $SE(T)$ to denote its standard error.
- In practice, standard errors are functions of the sample size and the original variability in the population from which we sampled our data.

# Confidence Intervals

- A confidence interval is a way of summarizing a sample statistic (e.g. sample mean) and its standard error at once.
- An (approximate) 95% confidence interval for the expectation (population mean), $\mu_X$, of a continuous random variable $X$ from a random sample $\{X_1, \ldots, X_n\}$, for large $n$, is

$$[\bar{X} - 2 \cdot SE(\bar{X}),\ \bar{X} + 2 \cdot SE(\bar{X})]$$

- Notice, this CI depends on the sample; i.e., it is a statistic.
- **Interpretation:** if we resample 100 times and calculate the 95% confidence interval for each new sample, then approximately 95 of those CIs will contain the true (unknown) population mean.

# Central Limit Theorem

## Central Limit Theorem (CLT)

Let $\{X_1, \ldots, X_n\}$ denote a random sample of $n$ independent observations from a common distribution with finite mean $\mu$ and finite variance $\sigma^2$. Recall the sample mean is given by

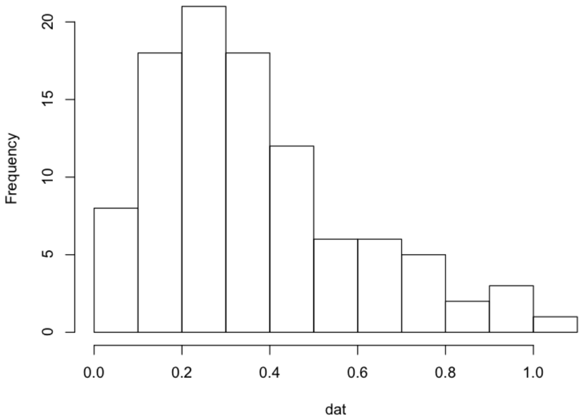
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, for $n$ large, $\bar{X}$ is approximately distributed as $N(\mu, \sigma^2/n)$.

- This is one of the most important theorems of classical statistics. Tells us all about how the sample mean behaves for an independent random sample from *any* common distribution with finite mean and variance.

# Central Limit Theorem: example

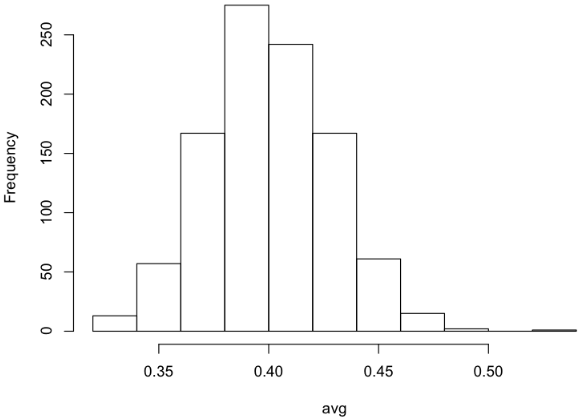
Histogram of **random sample** of size 100 from a very skewed (Gamma) random variable.




Sample mean is 0.366 for this particular set of 100 sample data points.

# Central Limit Theorem: example continued

Histogram of the **sample means** of 1000 random samples (each of size 100) from the same very skewed (Gamma) random variable.




Notice the histogram looks quite Normal! (CLT at work)

# Central Limit Theorem

- **Moral:** CLT allows us to treat the **sample mean** of *any* random phenomenon as a normal random variable, *as long as our sample size is big enough*.
- This will allow us to assign a measure of uncertainty to our sample mean estimate, e.g. by constructing *confidence intervals*.
- For small sample sizes, either the random phenomenon itself must follow a normal distribution, or we need to use other (nonparametric) statistical methods.

# Statistical Hypothesis Testing

- Nearly all quantitative science is based around the idea of stating and testing quantifiable hypotheses about study objects of interest.
- Point Null Hypothesis Testing (PNHT) is the most common option in virtually all applied disciplines.

# Statistical Hypothesis Testing

Basic recipe of PNHT:

(1) Identify parameter of interest.

(2) Define null hypothesis, $H_0$, of *no effect*.

(3) Define a *test statistic* $T$ (a function of the data) such that the larger $T$ is, the less consistent our data are with $H_0$.

(4) Collect data and then compute test statistic: $t_{obs}$.

(5) Compute p-value $= \Pr(|T| \geqslant t_{obs} \mid H_0)$; if p-value small enough, then conclude data are **inconsistent** with $H_0$.

Example:

(1) Difference in mean response between treatment groups $X$ and $Y$

(2) $H_0 : \mu_X = \mu_Y$

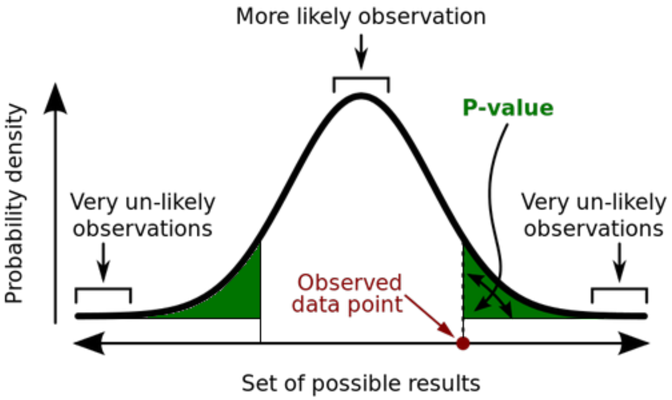
(3) $T$ = standardized difference in sample means

(4) Collect data; compute $t_{obs} = (\bar{X} - \bar{Y})/SE$

(5) Calculating p-value requires knowing distribution of $T$ given $H_0$....

# A Closer Look at P-values

- Under $H_0$, our test statistic follows some distribution (plotted).
- The **p-value** is the area under the test statistic's PDF (or PMF) that is *more extreme* than the observed test statisic from the sample.



# A Closer Look at P-values

- Formally, we define

$$\begin{aligned}\text{p-value} &= \Pr(\text{test stat. as or more extreme than observed} \mid H_0 \text{ true}) \\ &= \Pr(T \geqslant t_{obs} \mid H_0), \text{ usually.}\end{aligned}$$

- But think about this: the *evidence* that we observe is captured in the value of $t_{obs}$.
- The hypothesis we want to make a decision about is $H_0$.
- Think about the Sally Clark case: we would typically evaluate evidence for a hypothesis as $\Pr(H_0 \mid t_{obs})$.

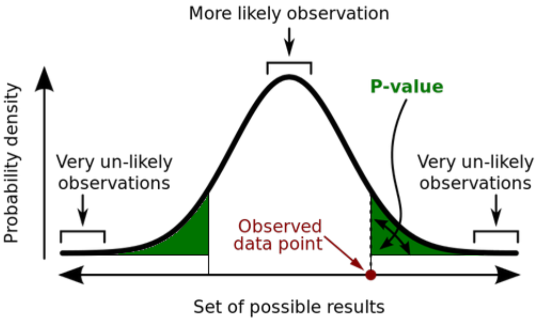
- But this is a very different conditional probability than what a p-value is!

# A Closer Look at P-values

- Formally, we define

$$\text{p-value} = \Pr(\,T \geqslant t_{obs} \mid H_0), \text{ usually.}$$

- **Interpretation:** the p-value is the probability of observing a test statistic as or more extreme than the one observed for our sample, given that the null hypothesis is true.



# A Closer Look at P-values

- Formally, we define

$$\text{p-value} = \Pr(T \geqslant t_{obs} \mid H_0), \text{ usually.}$$

- **Interpretation:** the p-value is the probability of observing a test statistic as or more extreme than the one observed for our sample, given that the null hypothesis is true.

- So a big p-value means the observed test statistic is "typical" under $H_0$. Therefore, the data are consistent with $H_0$.
- A small p-value means the observed test statistic is *not* "typical" under $H_0$. Therefore, the data are inconsistent with $H_0$.

# A Closer Look at P-values

- Formally, we define

$$\text{p-value} = \Pr(T \geqslant t_{obs} \mid H_0), \text{ usually.}$$

- Recall definition of conditional probability:

$$\Pr(T \geqslant t_{obs} \mid H_0 \text{ true}) = \frac{\Pr(T \geqslant t_{obs}, \text{ and } H_0 \text{ true})}{\Pr(H_0 \text{ true})}.$$

- With this in mind, how could the p-value be small?