Statistical inference: let X be a random variable from a distribution F. We don't know F. Given that we observe an outcome, denoted $X = x$, can we infer (some property of) F?

Common scenario: we have $X_1, \ldots, X_n$ random variables, all assumed to be *independent* and from distribution F. This is called a *random sample* of size n from F. The unknown F is typically called the *population*.

Abuse of notation: we will use X to mean a single random variable X or perhaps a random sample $X = (X_1, \ldots, X_n)$. The meaning of X will be clear from context.

We often assume F is a *parametric* distribution with some parameter θ. Example: each X_i is coin flip (1 if heads, 0 if tails) and $P(X_i = 1) = \theta$.

Hypotheses: we start with a theory about F, called the *null hypothesis*, denoted H_0, and ask if data provides enough evidence to reject the null hypothesis. If not, we might "accept" (or fail to reject) the null. Sometimes we explicitly specify an alternative hypothesis H_A (by default H_A is simply that H_0 is not true).

Hypotheses about parameters: A common hypothesis is one about an unknown parameter θ. First, we *assume* that F is a distribution parameterized by θ for some $\theta \in \Theta$. This assumption is *not* being tested! The null hypothesis H_0 is that $\theta \in \Theta_0$ and the alternative hypothesis is that it's not.

Example (coin): null hypothesis is that the coin is fair or almost fair: $\Theta_0 = [0.45, 0.55]$.

Hypothesis test: a function of random variable X that returns either "reject" or "accept". A typical form is a *threshold test*: choose some *test statistic* T and some constant c and then reject if $T(X) > c$, otherwise accept.

Example (coin): suppose we observe $n = 100$ coin flips, $X = (X_1, \ldots, X_n)$. Our test statistic is $T(X) = \left| \frac{\sum_{i=1} X_i}{n} - \frac{1}{2} \right|$ and we reject if $T(X) > 0.10$.

Errors: A *type I error* (false positive), reject H_0 when it's true; a *type II error* (false negative) accept H_0 when it's false.

Probability of screwing up: What is P(test says "reject")? It depends! It depends on whether H_0 is true or not. In the case of a parametric distribution, it depends on the (unknown) value of θ!

Let P_{H_0}(test says "reject") denote the probability that we reject the H_0 when it is true. The ideal test would have P_{H_0}(test says "reject") $= 0$ and P_{H_A}(test says "reject") $= 1$. Unfortunately, such tests do not exist in most situations.

Probability of error for tests of parameters In the case of a hypothesis tests about parameters, let P_θ(test says "reject") denote the probability that we reject the null hypothesis when the (unknown) parameter is equal to θ. When $\theta \in \Theta_0$, this is the probability of a Type I error; when $\theta \notin \Theta_0$, this is one minus the probability of a Type II error.

(This probability is a function of θ; this function is referred to as the *power function*.)

Example (coin): we can use simulation to estimate this probability. (Or, we can use math

and the fact that the coin flips follow a binomial distribution, which is closely approximated by a normal distribution – see textbook.)

The α level of a test: Consider this quantity: $\alpha = \max_{\theta \in \Theta_0} P_\theta(\text{test says "reject"})$ This is describing a worst-case scenario: if we choose the worst θ from the null hypothesis, what is the probability we reject the null? Tests are typically designed so $\alpha = 0.05$.

Let's assume we have a test of the form: if $T(X) > c$, then reject, otherwise accept. Then we can "design" the test by finding the c such that $0.05 = \max_{\theta \in \Theta_0} P_\theta(T(X) > c)$.

Example (coin): for the coin example, how do we set c so $\alpha = 0.05$? Step 1: figure out which $\theta \in \Theta_0$ is the worst-case. Step 2: figure out a c such that the probability of rejection is 0.05 (or close to it).

p-values: The choice of $\alpha = 0.05$ is pretty arbitrary; p-values offer an alternative approach. Suppose we've already observed our sample; in notation, we observe $X = x$ and therefore the observed value of our test statistic is $T(x)$. We can ask the following hypothetical, suppose I designed my test so that $T(x)$ was the cutoff, what is the worst-case probability of a Type I error? Formally, $p = \max_{\theta \in \Theta_0} P_\theta(T(X) \geq T(x))$.

Testing of multiple hypotheses: Suppose we have 10 different coins and we want to test whether they are fair. We employ the approach described above with $\alpha = 0.05$. Suppose the null hypotheses are all true: all of the coins are fair. What is the probability that we reject the null hypothesis for at least one of them? $P(\text{at least one rejection}) = 1 - P(\text{no rejections}) = 1 - (1 - \alpha)^{10} = 1 - (0.95)^{10} \approx 0.40$.

Moral of the story: α controls the error rate of a *single* hypothesis test. If you simultanenously test many hypotheses, one of them is bound to get rejected *by chance alone*. In other words, if you look hard enough, you're bound to find some interesting pattern in your data, even though that pattern may very well be due to chance.

There are ways to correct for testing of multiple hypothesis. One simple correction is the Bonferroni correction: if you test m hypotheses, then reject a null hypothesis if the p-value is $\leq \frac{\alpha}{m}$.

Example (coin): Again, assuming all 10 coins are fair, what is the probability that we reject the null hypothesis for at least one of them when we use the Bonferroni correction? $P(\text{at least one rejection}) = 1 - P(\text{no rejections}) = 1 - (1 - \frac{\alpha}{10})^{10} = 1 - (0.995)^{10} \approx 0.049$.

The practice of trying many hypotheses until you find one to reject is called p-hacking: `https://xkcd.com/882/`.

Conclusion There is an art and science to designing a good test for a given problem. It's important to have a good conceptual understanding of what hypothesis testing is. For a given situation, you can often look up the appropriate test.