Review of Basic Statistical Concepts

March 8, 2017

What is statistics?

The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.

- Descriptive Statistics
  Methods that organize and summarize data.
  - Numerical summary
  - Graphical methods
- Inferential Statistics
  Generalizing from a sample to the population from which it was selected.
  - Estimation
  - Hypothesis testing

- Population
  The entire collection of individuals or objects about which information is desired.
- Sample
  A subset of the population selected in some prescribed manner for study.

Numerical summaries

- Measure of central tendencies
  Mean, Median
- Measure of variability
  Variance,Standard deviation, Quartiles

- The Mean
  To find the mean of a set of observations, add their values and divide by the number of observations. If the n observations are $X_1, X_2, \cdots, X_n$, their mean is $\bar{X}$.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

  In a more compact notation,

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

- Median M
  The Median M is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger. To find the median of a distribution:
  1. Arrange all observations in order of size, from smallest to largest.
  2. If the number of observations n is odd, the median M is the center observation in the ordered list.
  3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list.

- Quartiles Q_1 and Q_3
  To calculate the quartiles:
  1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
  2. The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
  3. The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

- The Five Number Summary and Box-Plot
  The five number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five number summary is:

$$\textit{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \textit{Maximum}$$

- A box-plot is a graph of the five number summary.
  A central box spans the quartiles.
  A line in the box marks the median.
  Lines extend from the box out to the smallest and largest observations.
- Box-plots are most useful for side-by-side comparison of several distributions.

- The Variance S^2
  The Variance S^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $X_1, X_2, \cdots, X_n$ is

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

  Or, more compactly,

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

- The Standard Deviation
  The standard deviation S is the square root of the variance S^2:

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

  Computational formula for variance

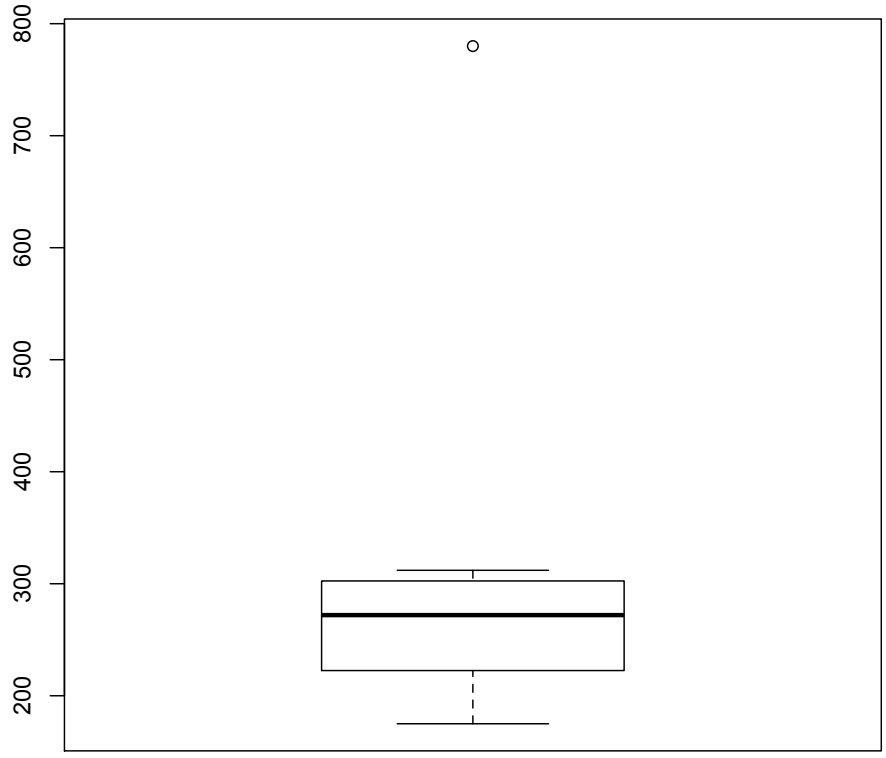
$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}$$

- Example 1: Books Page Length
  A sample of $n = 8$ books is selected from a librarys collection, and page length of each one is determined, resulting in the following data set.

$$X_1 = 247, X_2 = 312, X_3 = 198, X_4 = 780,$$

$$X_5 = 175, X_6 = 286, X_7 = 293, X_8 = 258$$

R code: X=c(247,312,198,780,175,286,293,258)
mean(X)=318.625, median(X)=272,
var(X)= 36945.12
sd(X)=192.2111
boxplot(X)



Introduction to Inference

- The purpose of inference is to draw conclusions from data.
- Conclusions take into account the natural variability in the data, therefore formal inference relies on probability to describe chance variation.
- We will go over the two most prominent types of formal statistical inference
  Confidence Intervals for estimating the value of a population parameter.
  Tests of significance which assess the evidence for a claim.
- Both types of inference are based on the sampling distribution of statistics.

In a typical statistical problem, we have a random variable X of interest, but its pdf $f(x)$ or pmf $p(x)$ is not known. Roughly, we can classify it into two ways:

- $f(x)$ or $p(x)$ is completely unkonwn (nonparametric inference);
- The form of $f(x)$ or $p(x)$ is known down to a parameter θ, where θ may be a vector (parametric inference).

Parameters and Statistics

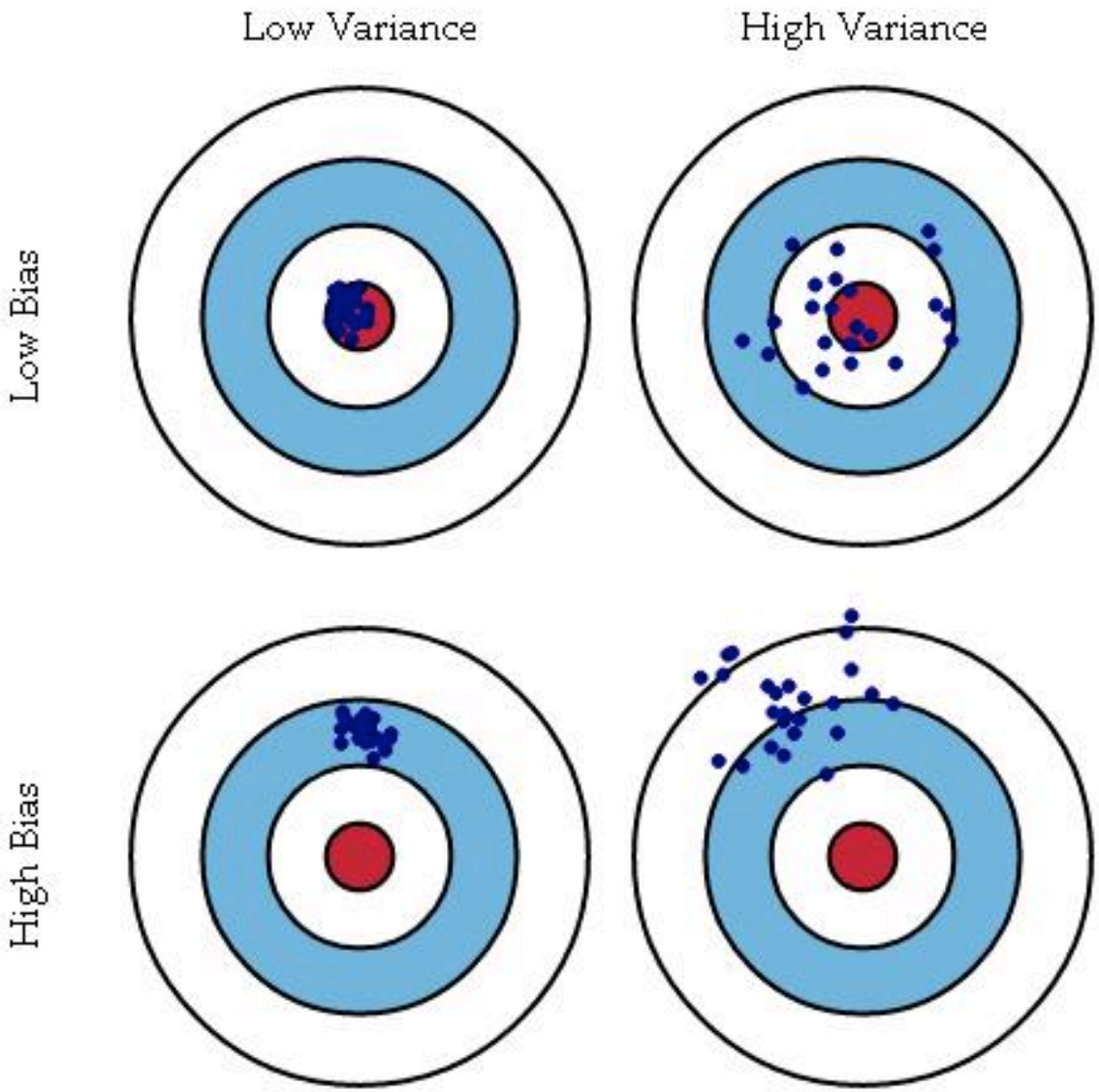
- A parameter is a number that describes the population.
  A parameter is a fixed number, but in practice we do not know its value.
- A statistic is a number that describes a sample.
  The value of a statistic is known when we have taken a sample, but it can change from sample to sample.
- We often use statistic to estimate an unknown parameter.

Example 2:Consumer attitude towards shopping

A recent survey asked a nationwide random sample of 2500 adults if they agreed or disagreed with the following statement:
I like buying new cloths, but shopping is often frustrating and time consuming.
Of the respondents, 1650 said they agreed.
The proportion of the sample who agreed that cloths shopping is often frustrating is:

$$\hat{P} = 1650/2500 = 0.66 = 66\%$$

- The number $\hat{P} = .66$ is a statistic.
- The corresponding parameter is the proportion(call it P) of all adults who would have said “Yes” if asked the same question.
- We don’t know the value of parameter P, so we use $\hat{P}$ as its estimate. That is, the value of $\hat{P}$ will vary from sample to sample.
- Random samples eliminate bias from the act of choosing a sample, but they can still be wrong because of the variability that results when we choose at random.


Low Variance
High Variance
Low Bias
High Bias

- The first advantage of choosing at random is that it eliminates bias.
- The second advantage is that if we take lots of random samples of the same size from the same population, the variation from sample to sample will follow a predictable pattern.
- **All statistical inference is based on one idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.**

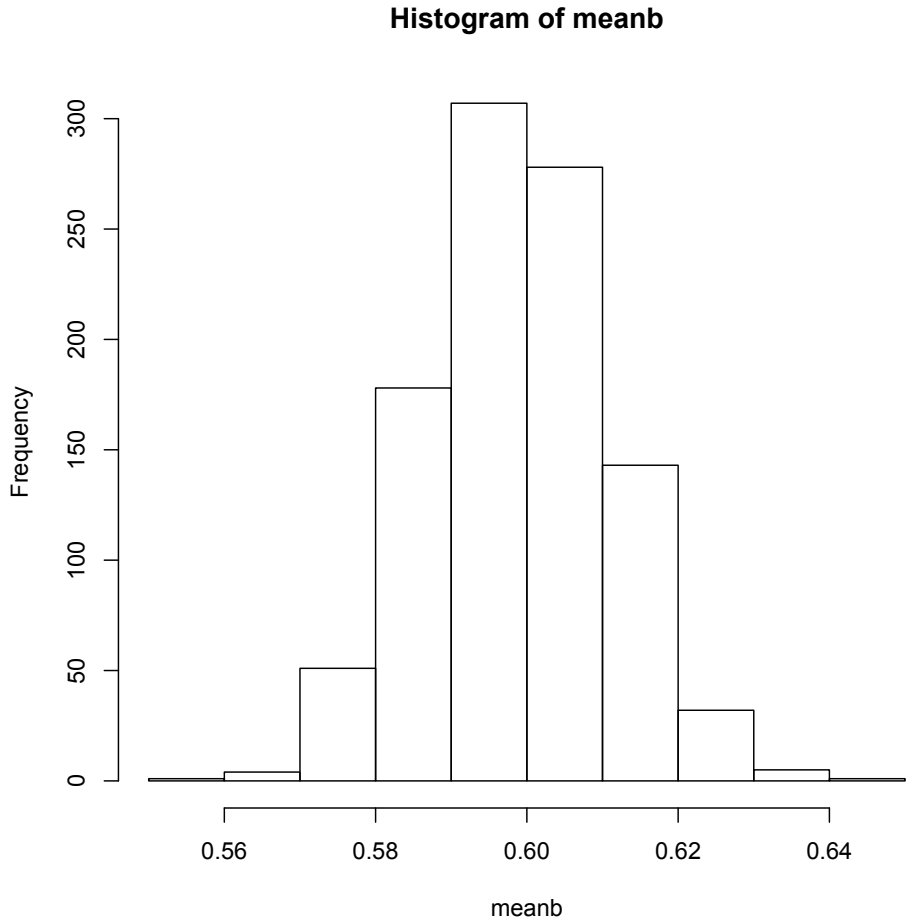
Sampling Distribution of Statistics

- Suppose that exactly 60% of adults find shopping for cloths frustrating and time consuming.
- That is, the truth about the population is that $P = 0.6$. (parameter)
- We select a simple random sample) of size 100 from this population and use the sample proportion($\hat{P}$, statistic) to estimate the unknown value of the population proportion P.
- What is the distribution of $\hat{P}$?

To answer this question:

- Take a large number of samples of size 100 from this population.
- Calculate the sample proportion $\hat{P}$ for each sample.
- Make a histogram of the values of $\hat{P}$.
- Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

A=rbinom(2500,1,0.6) sampleA=sample(A) meanb=mean(sampleA) hist(meanb)



Sampling Distribution

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

- Point Estimator
- Confidence sets
- hypothesis tests

Point estimator

Estimation of the parameters: Method of moments, Maxmum likelihood estimator

Method of moments

- Definition: Let X be a random variable. If EX^k is finite, where k is a positive integer, then EX^k is called the kth moment of X.
- By setting the kth moment of the population to be equal to the kth moment of the sample, we can get the estimation of the parameter.

- Example: Suppose the mean and variance of the random variable are μ and σ^2. $X_1, X_2, \cdots, X_n$ is a sample of X, estimate μ and σ^2. For X, we have

$$\alpha_1 = E(X) = \mu, \alpha_2 = E(X^2) = var(X) + [E(X)]^2 = \sigma^2 + \mu^2$$

For the sample, we have

$$A_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, A_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Then we have

$$\mu = \bar{X}, \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

By solving the equations, we have

$$\hat{\mu} = \bar{X}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- $\hat{\sigma}^2 = \frac{n-1}{n} S^2$.
- We have used our sample estimates as point estimates of parameters, for example: $\bar{X}$ as an estimate of μ, S^2 as an estimate of σ^2.
- **Why we use S^2 instead of $\hat{\sigma}^2$ as the estimate of σ^2??**

- $\bar{X}$ and S^2 are both unbiased estimators.
  The expected value of an unbiased estimator is equal to the parameter that it is trying to estimate.
- $\bar{X}$ is also a minimum variance estimator of μ.
  This means that it has the smallest variability among all estimators of μ.
- We will introduce Maximum Likelihood Estimator (MLE) for point estimator later on.

$$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

$$Var(\bar{x}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{1}{n}\sigma^2$$

$$\begin{aligned}
E(s^2) &= E(\frac{1}{n-1}\sum_{i}(x_i - \bar{x})^2) \\
&= \frac{1}{n-1}E(\sum_{i=1}^{n}((x_i - \mu) - (\bar{x} - \mu))^2 \\
&= \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - \mu)^2 - 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) + n(\bar{x} - \mu)^2) \\
&= \frac{1}{n-1}(n\sigma^2 - n \cdot \frac{1}{n}\sigma^2) = \sigma^2
\end{aligned}$$

What if we want to do more than just provide a point estimate? Are we confident with the point estimator obtained from the samples?

- Go back to Example 2. The estimator $\hat{P}$ will have different values if we take different samples.
- Suppose we can estimate this parameter from sample data, and we know the distribution of this estimator, then we can use this knowledge and construct a probability statement involving both the estimator and the true value of the parameter.
- This statement is manipulated mathematically to produce **confidence intervals**.

Central Limit Theorem

Let $X_1, X_2, \cdots, X_n$ denote the observations of a random sample from a distribution that has mean μ and positive variance σ^2. Then the random variable $Y_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges in distribution to a random variable which has a normal distribution with mean zero and variance 1.

Example: Estimating with Confidence

Community banks are banks with less than a billion dollars of assets. There are approximately 7500 such banks in the United States. In many studies of the industry these banks are considered separately from banks that have more than a billion dollars of assets. The latter banks are called “large institutions.” The community bankers Council of the American bankers Association (ABA) conducts an annual survey of community banks. For the 110 banks that make up the sample in a recent survey, the mean assets are $\bar{X} = 220$ (in millions of dollars). Suppose that the true standard deviation for all the community banks is equal to 161. What can we say about μ, the mean assets of all community banks?

- The sample mean $\bar{X}$ is the natural estimator of the unknown population mean.
- We know that
  $\bar{X}$ is an unbiased estimator of μ.
  The law of large numbers says that the sample mean must approach the population mean as the size of the sample grows.
- Therefore, the value X = 220 appears to be a reasonable estimate of the mean assets for all community banks.
- But, how reliable is this estimate?

Confidence intervals

Let X be a random variable with a probability distribution having statistical parameters θ, which is a quantity to be estimated. Suppose we have the samples $X_1, X_2, \cdots, X_n$, given the confidence level or confidence coefficient $1 - \alpha$, if there exist two statistics:$\hat{\theta}_1(X_1, X_2, \cdots, X_n)$ and $\hat{\theta}_2(X_1, X_2, \cdots, X_n)$, such that

$$P(\hat{\theta}_1(X_1, X_2, \cdots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \cdots, X_n)) = 1 - \alpha,$$

then $(\hat{\theta}_1, \hat{\theta}_2)$ is called the confidence interval.

Confidence Interval for the Population Mean

- We use the sampling distribution of the sample mean $\bar{X}$ to construct a level $1 - \alpha$ confidence interval for the mean μ of a population.
- We assume that data are a simple random sample of size n.
- The sampling distribution is exactly $N(\mu, \frac{\sigma^2}{n})$ when the population has the $N(\mu, \sigma^2)$ distribution.

- When σ^2 is known, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, then

$$P\{|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}| \leq Z_{\alpha/2}\} = 1-\alpha$$

The confidence interval is:

$$[\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

- When σ^2 is unknown,

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1),$$

then

$$P\{|\frac{\bar{X}-\mu}{S/\sqrt{n}}| \leq t_{\alpha/2}\} = 1-\alpha$$

The confidence interval is:

$$[\bar{X} - t_{\alpha/2}(n-1)\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{\sigma}{\sqrt{n}}]$$

Standard Error of Estimator

- By Central Limit theorem, in repeated sampling the sample mean $\bar{X}$ is approximately normal, centered at the unknown population mean μ,with standard deviation

$$\bar{\sigma}_x = \frac{161}{\sqrt{110}} = 15$$

millions of dollars.

Confidence Interval for a Population Mean

- What is a 90% confidence interval for the mean assets of all community banks?
- $P(X < 1.65) = 0.95$

Tests of Significance

- Confidence intervals are appropriate when our goal is to estimate a population parameter.
- The second type of inference is directed at assessing the evidence provided by the data in favor of some claim about the population.
- A significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess.
- The hypothesis is a statement about the parameters in a population or model.
- The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree.

Statistical Tests

- How to tell if something is different from something else
- Test your intuition:
  1. Under what circumstances does the mean of a sample equal the mean of the population from which it was drawn?
  2. What about the standard deviation?
  3. What if your sample was very small relative to the population?

Statistical Tests

- Consider a simple example:
  You are testing the hypothesis that eating walnuts makes people smarter by feeding walnuts to a group of 30 subjects and then testing their IQ. We have known $\mu = 100, \sigma = 15$ for all the people.
- If you are right, then eating walnuts will make the average IQ of your subjects be higher than the average IQ of all people (the population) since, mostly, those other people don't eat walnuts much.

- Put another way:
  Is this sample(entirely) of walnut eaters different from the population of mostly non-walnut-eaters?

Types of Errors

There are two “mistakes” you could make:

- Type I error or False-Positive: you decide the walnut treatment works when it doesn’t really
- Type II error or False-Negative: you decide the walnuts don’t work when really they do

Types of Successes

There are two ways to succeed:

- Hit or True-Positive: You decide the walnuts do make people smarter and, in fact, they really do
- Correct-Rejection or True-Negative: You decide the walnuts don't work and, in fact they really don't

Statistical Tests

- Your subjects turn out to have a mean IQ of 107.5 (1/2 S.D. from the mean of the population) after eating walnuts
- What are two reasons why the mean IQ of your subjects might be greater than the mean of the population?
  1. You happened to pick 30 very smart people (i.e. university students)
  WARNING: Type I error is possible!
  2. The walnuts worked.

- Usually we are worried about making a type I error so we need to know:
  What fraction of all possible groups of 30 subjects would have a mean IQ of 105 or less?
- In other words, we are interested not in the distribution of IQ scores themselves, but rather in the distribution of mean IQ scores for groups of 30 subjects

Example Z-Test

Using our example in which we are testing the hypothesis that walnuts make people smarter
Null hypothesis is that they don't

$$\bar{X} = 107.5, \mu = 100, \sigma = 15$$

Here's what we've got:

$$\bar{X} = 107.5, \mu = \mu_{\bar{X}} = 100, \sigma = 15, n = 30$$

We can compute:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{30}} = 2.738$$

Looking up 2.739 in the Z table reveals that only .0031 or .31% of the means in the sampling distribution of mean IQs (for groups of 30 people each) would have a mean equal to or greater than 107.5!

- What this means is that you have only a 0.31% chance of making a type I error if you conclude that walnuts made your subjects smarter!
- Put another way, there is only a 0.31% chance that this sample of IQs is taken from the regular population...walnut eaters are different!

Alpha

- Is .31% small enough? What risk of making a Type I error is too great?
- There is no absolute answer - it depends entirely on the circumstances
- 5% or probability p= .05 is generally accepted
- This rate of making Type I errors (ie. number of Type I errors per 100 experiments) is called the Alpha Level

Tests of Significance: Formal details

- The first step in a test of significance is to state a claim that we will try to find evidence against.
- Null Hypothesis H_0
  1. The statement being tested in a test of significance is called the null hypothesis.
  2. The test of significance is designed to assess the strength of the evidence against the null hypothesis.
  3. Usually the null hypothesis is a statement of "no effect" or "no difference." We abbreviate "null hypothesis" as H_0.

Tests of Significance: Formal details

- A null hypothesis is a statement about a population, expressed in terms of some parameter or parameters.
- The null hypothesis in our walnut example is $H_0 : \mu \leq 100$
- It is convenient also to give a name to the statement we hope or suspect is true instead of H_0.
- This is called the alternative hypothesis and is abbreviated as H_a or H_1.
- In our walnut example the alternative hypothesis states that average IQ after eating walnut is higher than 100. We write this as $H_1 : \mu > 100$

- Since H_1 expresses the effect that we hope to find evidence for we often begin with H_1 and then set up H_0 as the statement that the Hoped-for effect is not present.
- Stating H_1 is not always straight forward.
- It is not always clear whether H_1 should be one-sided or two-sided.

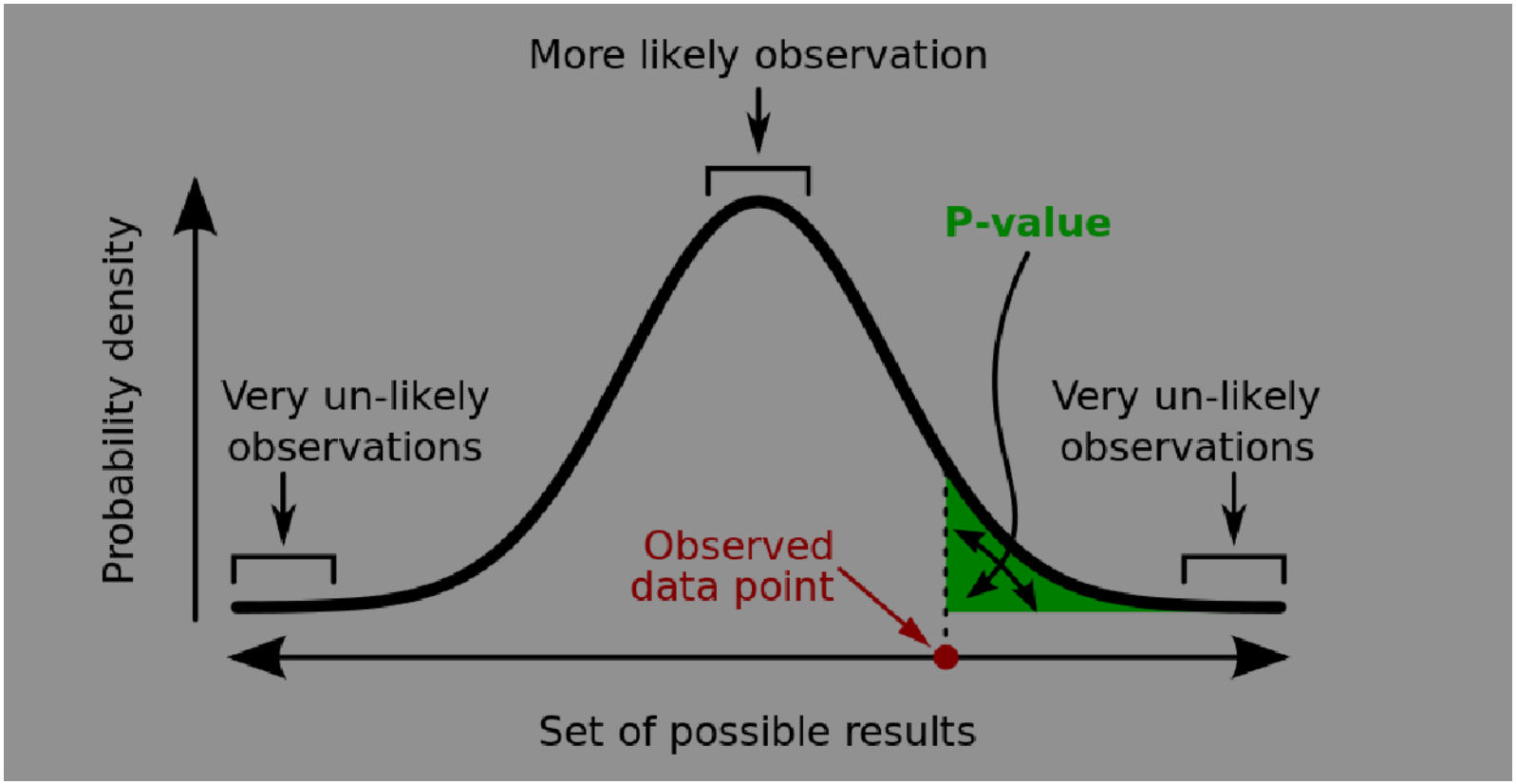
Test statistics

We will learn the form of significance tests in a number of common situations. Here are some principles that apply to most tests and that help in understanding the form of tests:

- The test is based on a statistic that estimate the parameter appearing in the hypotheses.
- Values of the estimate far from the parameter value specified by H_0 gives evidence against H_0.

P-value

- p-value is the probability that the test statistic would take a value as large or larger than one observed assuming that H_0 is true.
- The p-value is given by
  $Pr(X \geq x|H_0)$ for right tail event,
  $Pr(X \leq x|H_0)$ for left tail event,
  $2\min\{Pr(X \leq x|H_0), Pr(X \geq x|H_0)\}$ for double tail event.
- The level that says "this evidence is strong enough" is called significance level and is denoted by letter α.
- We compare the p-value with the significance level.
- We reject H_0 if the p-value is smaller than the significance level, and say that the data are statistically significant at level α.


More likely observation
P-value
Probability density
Very un-likely observations
Very un-likely observations
Observed data point
Set of possible results

Example: Coin flipping

Suppose that the experimental results show the coin turning up heads 14 times out of 20 total flips. The null hypothesis is that the coin is **fair**, and the test statistic is the number of heads. If we consider a right-tailed test, the p-value of this result is the chance of a fair coin landing on heads at least 14 times out of 20 flips. This probability can be computed from binomial coefficients as

$$\begin{aligned}&\text{Prob}(14 \text{ heads}) + \text{Prob}(15 \text{ heads}) + \cdots + \text{Prob}(20 \text{ heads}) \\ &= \frac{1}{2^{20}}\left[\binom{20}{14} + \binom{20}{15} + \cdots + \binom{20}{20}\right] = \frac{60{,}460}{1{,}048{,}576} \approx 0.058\end{aligned}$$

- Null hypothesis (H_0): The coin is fair, i.e. Prob(heads) = 0.5
- Test statistic: Number of heads
- Level of significance: 0.05
- Observation O: 14 heads out of 20 flips; and
- Two-tailed p-value of observation O given H_0 = 2*min(Prob(no. of heads$\geq$ 14 heads), Prob(no. of heads$\leq$ 14 heads))= 2*min(0.058, 0.978) = 2*0.058 = 0.115.
- p-value$>$ 0.05, we accept H_0.