Elements of Statistical Methods
Inference (Ch 9)

Fritz Scholz

Spring Quarter 2010

May 7, 2010
Inference as Inverse to Probability Calculations

When $Y \sim \text{Binomial}(n = 100, p = 0.5)$ we can calculate $P(40 \leq Y \leq 60)$ as

$$P(Y \leq 60) - P(Y \leq 39) = \text{pbinom}(60, 100, .5) - \text{pbinom}(39, 100, .5) = 0.9647998$$

We can do the same for any other specified value $p$, just change the .5 above.

Inference addresses the reverse question. The parameter $p$ is unknown.

Based on the observed $Y = y$, what can we say about the unknown $p$?

In this course we discuss three different modes of inference about $p$.

Point estimation, hypothesis testing and confidence sets
1-Sample Problem: Inference About $\mu$

The previous binomial example is a special case of a more general situation.

We will first address all inference problems in the context of the 1-sample problem described as follows

1. $X_1, \ldots, X_n$ i.i.d. $\sim F$. We observe a sample $\bar{x} = \{x_1, \ldots, x_n\}$.

2. Both $EX_i = \mu$ and $\text{var} X_i = \sigma^2$ exist and are finite.

   Draw inferences about the population mean $\mu$, which is fixed but unknown.
   In the binomial case we have $\mu = p$ and $\sigma^2 = p(1 - p)$.

3. The sample size $n$ is sufficiently large so that we can use the normal approximation provided by the CLT.

   This widens the applicability scope w.r.t. $F$. 
Binomial Example

Suppose someone proposes to make a fair chance decision by spinning a coin. Being somewhat suspicious about this, we carry out $n = 100$ spins of the coin and observe $Y = y = 32$ Heads.

We can ask the following questions about this coin spinning process.

1. What is a reasonable guess as to the true value of $p = P(\text{Heads})$.

2. Is the position of fairness, i.e., $p = 0.5$, believable or should we reject it?

3. What kind of values $p$ are plausible in view of the observed $y = 32$?
For Bernoulli r.v.’s we have $EX_i = \mu = p$ and $Y = X_1 + \ldots + X_n \sim \text{Binomial}(n, p)$.

The law of large numbers $\implies \hat{F}_n(y) \xrightarrow{P} F(y)$, combined with the plug-in principle, i.e.,

use $\bar{X}_n$ as the mean of $\hat{F}_n$ in parallel to $\mu$ as the mean of $F$, with $\bar{X}_n \xrightarrow{P} \mu$,

suggests the use of the point estimate

$$\hat{p} = \bar{x}_n = \frac{y}{n} = \frac{32}{100} = 0.32$$

“point” refers to the fact that a single number, a point on the number line, is reported.
Hypothesis Testing

Since coin spinning was claimed to be a fair process, is this position still defensible in view of $\bar{x}_n = 0.32$, i.e., $y = 32$?

Suppose $p = 0.5$ is true, how likely is it to observe a proportion $\bar{X}_n$ that differs from 0.5 by as much as $|0.5 - 0.32| = 0.18$ or more, i.e., to observe a $Y$ that differs from $n \cdot p = 100 \cdot 0.5 = 50$ by 18 or more?

\[
p = P(|Y - 50| \geq 18) = P(Y \leq 32 \cup Y \geq 68) = P(Y \leq 32) + P(Y \geq 68)
\]
\[
= P(Y \leq 32) + 1 - P(Y \leq 67)
\]
\[
= \text{pbinom}(32, 100, .5) + 1 - \text{pbinom}(67, 100, .5) = 0.0004087772
\]

This significance probability or $p$-value $p$ is so small that the supposition or hypothesis $p = 0.5$ and chance alone do not provide a believable explanation.

At what small value $p$ does the hypothesis become unacceptable? The appropriate choice depends on the circumstances (more later).
Set Estimation or Confidence Sets

Here we relax the requirement of giving a single number as a point estimate. We ask for a range of values for $p$ that appear to be plausible or acceptable.

The previous calculation of $p$ hypothesized $p_0 = 0.5$ as the true value for $p$. To make this dependence on $p_0$ and $y$ more explicit we also write $p(y; p_0)$.

We can test other values $p_0 \in [0, 1]$ as possible hypotheses for the value of $p$, each time obtaining a value $p(y; p_0)$ (just replace 0.5 by $p_0$ in the previous calculation).

When $p(y; p_0)$ is sufficiently small, say $p(y; p_0) \leq 0.1$ (or $\leq 0.05$ or $\leq 0.01$) we declare such values $p_0$ as not acceptable or not plausible.

However, any $p_0$ with $p(y; p_0) > 0.1$ would be judged plausible. The set of such plausible $p_0$ is our set estimate of $p$, also called a 90\% (or 95\% or 99\%) confidence set for the unknown $p$. 
Calculation of Confidence Sets

Calculate the following for a fine grid of values for \( p_0 \)

\[
\begin{align*}
p(y = 32; p_0) &= P_{p_0}(|Y - np_0| \geq |32 - np_0|) \\
&= P_{p_0}(\{Y - np_0 \geq |32 - np_0|\} \cup \{Y - np_0 \leq -|32 - np_0|\}) \\
&= * \quad P_{p_0}(Y - np_0 \geq |32 - np_0|) + P_{p_0}(Y - np_0 \leq -|32 - np_0|) \\
&= P_{p_0}(Y \geq np_0 + |32 - np_0|) + P_{p_0}(Y \leq np_0 - |32 - np_0|) \\
&= 1 - \text{pbinom}(\text{ceiling}(n \times p_0 + \text{abs}(32 - n \times p_0)) - 1, n, p_0) \\
&\quad + \text{pbinom}(n \times p_0 - \text{abs}(32 - n \times p_0), n, p_0)
\end{align*}
\]

and find the set (interval) of values \( p_0 \) where \( p(y = 32; p_0) > 0.1 \).

\(^*\) is an \( = \) when there is no overlap in the union. When there is an overlap, e.g., when \( 32 = np_0 \), the value of \( p(y = 32; p_0) \) is 1.
pvalBinom.plot<-function(y=32,n=100,alpha=.1){
  p <- seq(.0001,.9999,.0001)
  Delta <- abs(y-n*p) # a vector of length 9999
  pval <- 1-pbinom(ceiling(n*p+Delta)-1,n,p)+pbinom(n*p-Delta,n,p)
  pval[pval>1] <- 1
  p1 <- min(p[pval>alpha]); p2 <- max(p[pval>alpha])
  plot(p,pval,ylab="p-value",pch=16,cex=.5)
  lines(p,pval)
  abline(v=c(p1,p2)); abline(h=alpha)
  text(p1-.01,.5,signif(p1,3),adj=1)
  text(p2+.01,.5,signif(p2,3),adj=0)
  text(1,.95,paste("n =",n," y =",y),adj=1)
}

# ceiling(x) is the smallest integer >= x (vectorized)

Note the vectorized calculation for all 9999 values of p at once.
Output from pvalBinom.plot

n = 100, y = 32

Note that the interval does not cover 0.5.
Increasing the Sample Size

\[ n = 1000, y = 320 \]
Coverage Probability

Let $C(y)$ be the set of all acceptable $p_0$, i.e., those for which $p(y; p_0) > 0.1$

$$C(y) = \{p_0 : p(y; p_0) > 0.1\}$$

When viewing $C(y)$ with $y$ replaced by the corresponding r.v. $Y$, it becomes a random set $C(Y)$ for which the following coverage probability statement holds

$$P_{p_0}(p_0 \in C(Y)) \geq 1 - 0.1 = 0.9$$

At least 90% of such random sets will cover the true but unknown target $p_0$, no matter what it is. It is not $p_0$ that is random here!

90% or 0.9 indicates the confidence level of such sets. Changing 0.1 to 0.05 or 0.01 yields higher confidence levels 95% or 99%.

However, for any given confidence set we will not know whether $p_0 \in C(Y)$ or not.
The Coverage Argument

\[ P_{p_0}(p_0 \in C(Y)) = 1 - P_{p_0}(p_0 \notin C(Y)) = 1 - P_{p_0} \left( Y \text{ is sufficiently extreme for rejecting } H_0 : p = p_0 \right) \geq 1 - 0.1 = 0.9 \]

since \( Y \) is sufficiently extreme when \( p(Y; p_0) \leq 0.1 \)

The argument is the same for any other number \( \alpha \in (0, 1) \) different from 0.1.

The coverage probability or confidence level then becomes \( 1 - \alpha \).
Point Estimation

The goal of point estimation is to make a reasonable guess of the unknown population parameter or characteristic of interest, e.g., of the population mean $\mu$.

This quantity to be estimated is also called the estimand.

When estimating $\mu$ and using the sample mean as estimate we distinguish between

$$
\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
$$

Here $\bar{x}_n$ is referred to as an estimate, based on the observed values $\vec{x} = \{x_1, \ldots, x_n\}$. $\bar{x}_n$ is a single number, no longer subject to chance variation.

$\bar{X}_n$ is referred to as an estimator, the procedure to use for all potential samples $\vec{X} = \{X_1, \ldots, X_n\}$. $\bar{X}_n$ is a random variable, i.e., subject to chance variation.

Any function of $\vec{X} = \{X_1, \ldots, X_n\}$ is also called a statistic.
Properties of Estimators

Usually, it is not possible to say how close an estimate is to its target. However, we can say something about the behavior of estimators.

For example, we know $E\bar{X}_n = \mu$. The mean of the $\bar{X}_n$ population coincides with the target $\mu$, the mean of the sampled population. We say that $\bar{X}_n$ is an unbiased estimator of $\mu$.

$X_4$ is also an unbiased estimator of $\mu$, since $EX_4 = \mu$.

However

$$\text{var} \bar{X}_n = \frac{\sigma^2}{n} \quad \text{while} \quad \text{var} X_4 = \sigma^2$$

Thus the scatter of the $\bar{X}_n$ is tighter by a factor of $1/\sqrt{n}$ than the scatter of $X_4$.

$$\bar{X}_n \xrightarrow{P} \mu \quad \text{as} \quad n \to \infty \quad \text{but} \quad X_4 \xrightarrow{P} \mu$$

We say that $\bar{X}_n$ is a consistent estimator of $\mu$.

Choosing inconsistent estimators makes little sense.
Estimators of $\sigma^2$

Another estimand of intrinsic interest is the population variance $\sigma^2$.

The plug-in principle suggested the estimate $\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 / n$.

The corresponding plug-in estimator is biased since

$$E \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right] = \frac{n-1}{n} \sigma^2 = \sigma^2 - \frac{\sigma^2}{n} < \sigma^2$$

However, the alternate estimator, called the sample variance,

$$S_n^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right] = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is unbiased, since

$$ES_n^2 = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2$$

Both estimators are reasonable and consistent.

The square root of either estimator is a biased estimator of $\sigma$. 
Section 9.3 gives a lengthy and interesting discussion of various aspects of testing hypotheses. There is a very strong parallel with criminal trials. (READ!)

Based on a random sample (or more generally data) from some population, say $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, decide whether the data arose from one subset of these populations or its complement, e.g., decide $p \in A \subset [0, 1]$ or $p \in A^c$.

In our binomial example we had $A = \{0.5\}$ and $A^c = \{p : p \neq 0.5\}$.

Usually these subsets are associated with hypotheses, the null hypothesis $H_0$ and the alternative hypothesis $H_1$.

In our binomial example these would be described as $H_0 : p = 0.5$ and $H_1 : p \neq 0.5$. 
Decision Theoretic Model

In a more generic setting we would have a family of possible probability models
\[ \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \], where each \( P_\theta \) could have given rise to the random sample (data) \( X_1, \ldots, X_n \). \( \theta \) can be a single real valued parameter or can be more complex.

Each particular probability model \( P_\theta \), or simply its indexing parameter \( \theta \), represents a state of nature. \( \Theta \) represents all such states under consideration.

Viewing \( \Theta \) as the disjoint union of \( \Theta_0 \) and \( \Theta_1 \), i.e., \( \Theta_0 \cap \Theta_1 = \emptyset \) and \( \Theta_0 \cup \Theta_1 = \Theta \), we make one of two possible decisions about \( \theta \): \( \theta \in \Theta_0 \) or \( \theta \in \Theta_1 \).

Any such decision based on \( X_1, \ldots, X_n \) constitutes a test of the hypothesis \( H_0 : \theta \in \Theta_0 \) against the alternative \( H_1 : \theta \in \Theta_1 \).

Since we don’t know the true state of nature giving rise to \( X_1, \ldots, X_n \) we can’t be certain about having made the correct decision.
Type I and Type II Error

Hypothesis testing can be viewed as a game the statistician plays against nature. Nature chooses a $\theta \in \Theta = \Theta_0 \cup \Theta_1$. Based on a sample $X_1, \ldots, X_n$ from the unknown $P_\theta$ the statistician chooses between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$.

<table>
<thead>
<tr>
<th>Decision Maker’s Choice</th>
<th>$H_0 : \theta \in \Theta_0$</th>
<th>$H_1 : \theta \in \Theta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>correct decision</td>
<td>Type II Error</td>
</tr>
<tr>
<td>$H_1$</td>
<td>Type I Error</td>
<td>correct decision</td>
</tr>
</tbody>
</table>

There are 4 possible outcomes to this game.

For $\theta \in \Theta_0$: $P_\theta(\text{decide } H_1) = \text{type I error probability}$. For $\theta \in \Theta_1$: $P_\theta(\text{decide } H_0) = \text{type II error probability.}$
Error Probability Trade-off

View this decision problem as designating a portion $\mathcal{R}$ in the space of all $(X_1, \ldots, X_n)$ values for which we reject $H_0$, while for any $(X_1, \ldots, X_n) \in \mathcal{R}^c$ we accept $H_0$.

$$P_\theta(\text{type I error}) = P_\theta((X_1, \ldots, X_n) \in \mathcal{R}) = P_\theta(\mathcal{R}) \quad \text{for } \theta \in \Theta_0$$

$$P_\theta(\text{type II error}) = P_\theta((X_1, \ldots, X_n) \in \mathcal{R}^c) = P_\theta(\mathcal{R}^c) \quad \text{for } \theta \in \Theta_1$$

We can reduce $P_\theta(\text{type I error}) = P_\theta(\mathcal{R})$ by making the rejection region $\mathcal{R}$ smaller.

However, this increases $\mathcal{R}^c$ and thus increases $P_\theta(\text{type II error}) = P_\theta(\mathcal{R}^c)$.

Similarly, decreasing $P_\theta(\text{type II error})$ increases $P_\theta(\text{type I error})$.

The only way to drive down both probabilities is to increase the sample size $n$. 
Neyman-Pearson Formulation of Hypothesis Testing

This trade-off problem in the two types of error probabilities was resolved by Neyman and Pearson by treating the $P_\theta$ (type I error) as special, namely by placing a limit $\alpha \in (0, 1)$ on it, and among all level $\alpha$ tests, i.e., rejection regions $\mathcal{R}$ with

$$P_\theta(\text{type I error}) = P_\theta(\mathcal{R}) \leq \alpha \quad \text{for } \theta \in \Theta_0$$

they suggested to find a region $\mathcal{R}$ for which the $P_\theta$ (type II error) = $P_\theta(\mathcal{R}_c)$ is as small as possible for $\theta \in \Theta_1$.

This amounts to the same thing as finding a level $\alpha$ rejection region $\mathcal{R}$ for which

$$P_\theta(\mathcal{R}) = 1 - P_\theta(\mathcal{R}_c)$$

is as large as possible when $\theta \in \Theta_1$.

This bound $\alpha$ is also referred to as significance level (not always achievable).

This formulation turned out to be very fruitful and led to tests that often were intuitively appealing and were already widely used.
The Choice of $\alpha$

$\alpha$ controls or limits the probability of type I error, i.e., the probability of wrongfully rejecting $H_0$ when it is true. We test at significance level $\alpha$.

Usually, we want that limit to be small, depending on the importance of the error.

Customary values are $\alpha = 0.10, 0.05, 0.02, 0.01, 0.001$ or smaller. Very entrenched are 0.05 and 0.01, but that is more a matter of habit (Tables).

Choose $\alpha$ small to be fairly sure to have made the right decision when rejecting $H_0$, because then the chance of having made a wrong rejection of $H_0$ is very small.

Choose $\alpha$ not so small when the type I error is not so serious and when you are willing to be more open to alternative hypothesis explanations of the data.
In DNA microarray data situations a very large number, say \(N\), of significance tests is very common.

If each test is performed at significance level \(\alpha\) and if in all these tests \(H_0\) is true, then one could expect about \(N\alpha\) false rejections of \(H_0\).

This might lead to a large number of wild goose chases when no effect is present.

To guard against this, use \(\alpha = \alpha^*/N\) as the individual test significance level, when aiming for an overall false alarm rate of \(\alpha^*\).  \(\implies N\alpha = \alpha^*\).
The Choice of $\Theta_0$ and $\Theta_1$

Since $\Theta_0$ plays a special role in the Neyman-Pearson formulation, the question arises which subset of $\Theta$ to designate as $\Theta_0$.

Let us revert back to the coin spinning example. With $\theta = p$ and $\Theta = [0, 1]$ we had $\Theta_0 = \{0.5\}$ and $\Theta_1 = \{p : p \neq 0.5\}$.

“Conventional wisdom” suggested $p = 0.5$, until “proven” otherwise by sufficient evidence in the form of $X_1, \ldots, X_n$.

Since we had doubts we hope to overturn conventional wisdom. We set up $\Theta_0 = \{0.5\}$ as null hypothesis, which we hope to reject.

By our asymmetric treatment of $\Theta_0$ and $\Theta_1$ we control the chance of wrongfully rejecting $H_0$ (overturn wisdom) by $\alpha$. 
Another view: This asymmetric treatment of $H_0$ and $H_1$ is of conservative nature. We will stick with a simple chance explanation (under $H_0$) of any perceived effects, unless those effects (the data evidence) are strong enough.

The criminal trials parallel: The accused is innocent until proven guilty. In dubio pro reo (Lex Romana). The jury will have to decide what is guilt beyond reasonable doubt. ($\alpha = ?$)

By not rejecting $H_0$ (innocence) we don’t necessarily accept $H_0$ as the truth. Some prefer: We find the accused “not guilty” (not the same as “innocent”).

When such nuances are understood, it is easier to say: accept $H_0$ or accept $H_1$.

Probability calculations under $H_0$ need to be easier, because of the $\alpha$ requirement. This may influence the choice of $H_0$. It may also provide an easier explanation.

If $p \neq 0.5$, but $p = 0.500001$, who would want to distinguish that from 0.5?
Test Statistic

Usually the rejection region $\mathcal{R}$ in the set of all possible $(X_1, \ldots, X_n)$ values is defined in terms of a test statistic, say $W = W(X_1, \ldots, X_n)$, e.g.,

$$
\mathcal{R} = \{(X_1, \ldots, X_n) : W(X_1, \ldots, X_n) \geq w\} = \{(X_1, \ldots, X_n) : W \geq w}\}
$$

Sometimes it is more appropriate to reject $H_0$ when $W \leq w$, i.e., for whatever are unexpected values of $W$ under $H_0$.

However, any test of one type can be transformed into one of the other type, simply by transitioning from $W$ to $\tilde{W} = -W$ as test statistic.

$$
\{\tilde{W} : \tilde{W} \leq \tilde{w}\} = \{-W : -W \leq \tilde{w}\} = \{W : W \geq -\tilde{w}\} = \{W : W \geq w\}
$$

with $w = -\tilde{w}$
Here $X_1, \ldots, X_n$, the Bernoulli r.v.s, could be our basic data, or $Y = X_1 + \ldots + X_n$ could be the data. The order of successes does not matter. When testing $H_0 : p = p_0 = 0.5$ against $H_1 : p \neq p_0$ for $n = 100$, we use $W = |Y - 50| = |Y - np_0|$ as test statistic, rejecting $H_0$ when $W$ is too large, say when $W \geq c$.

For a level $\alpha$ test we need to satisfy the condition $P_{p_0}(W \geq c) \leq \alpha$. Subject to this condition, make the rejection region $R_c$ as large as possible, i.e., $R^c_c$ as small as possible, to minimize $P_p(\text{type II error}) = P_p(R^c_c)$ for $p \neq p_0$. Thus let $c_\alpha$ be the smallest value of $c$ for which $P_{p_0}(W \geq c) \leq \alpha$. Then we have $P_{p_0}(W \geq c_\alpha) \leq \alpha$. $c_\alpha$ is called the critical value of the test.
Connection to Significance Probability

In our previous treatment of this example we calculated the significance probability

\[ P_{p_0}(W \geq w) = P_{p_0}(|Y - np_0| \geq |y - np_0|) = p(y; p_0) \]

for the observed value \( w = |y - np_0| = |32 - 50| \) of \( W = |Y - np_0| = |Y - 50| \).

A very small value of \( p(y; p_0) \) was interpreted as very strong evidence against \( H_0 \), i.e., would give us a strong reason to reject \( H_0 \).

Note the following equivalence in relation to the critical value \( c_\alpha \)

\[ P_{p_0}(W \geq w) = p(y; p_0) \leq \alpha \iff w \geq c_\alpha \]

By definition, \( c_\alpha \) is the smallest \( w \) satisfying the inequality on the left side.

Thus \( p(y; p_0) \leq \alpha \) is equivalent to rejecting \( H_0 \) at significance level \( \alpha \).
Significance Probability & Rejection

\[ P(W \geq c) = P(Y - np_0 \geq c) \]

\[ n = 100 \]
\[ p_0 = 0.5 \]
\[ y = 38 \]
\[ w = |38 - 100 \times 0.5| \]
\[ p(y; p_0) = P(W \geq w) \]
\[ = 0.021 \]

\[ \alpha \Rightarrow w \geq c_\alpha \]
\[ \Rightarrow \text{reject } H_0 : p = p_0 \]

\[ \alpha = 0.05 \]
\[ \alpha_{\text{achieved}} = 0.0352 \]
A Priori Choice of $\alpha$

The significance level $\alpha$ should be chosen prior to observing any data.

This prevents after the fact manipulation of the decision outcome, i.e., tipping the scale.

The a priori choice of $\alpha$ defines a class of test procedures (all level $\alpha$ tests) among which the best is chosen according to the Neyman-Pearson theory.

On the other hand, if we reject $H_0$ at level $\alpha$, we don’t really know how strong or marginal the rejection was. Would we still have rejected $H_0$ at a much smaller $\alpha$?

The significance probability or $p$-value $p(y; p_0)$ captures this aspect much better. It is the preferred mode of reporting test results.

It gives anybody the option to decide with their a priori choice of $\alpha$. (slippery!)
Suppose we have a sample $X_1, \ldots, X_n$ from some population with mean $\mu$.

Some scientific theory may suggest a specific value $\mu_0$ for $\mu$.

Thus we might want to test the hypothesis $H_0 : \mu = \mu_0$ based on $X_1, \ldots, X_n$.

We have $\bar{X}_n$ as an estimator of the unknown $\mu$, whatever it is.

For large $n$ (assumed for now) we also know that it is close to $\mu$ (consistency).

$|\bar{X}_n - \mu_0|$ should give us some indication about $|\mu - \mu_0|$, i.e.,
is it sufficiently different from zero to conclude $|\mu - \mu_0| > 0$ or $\mu \neq \mu_0$?

Can we evaluate the significance probability for an observed value $\bar{x}_n$ of $\bar{X}_n$, i.e.,

$$p(\bar{x}_n; \mu_0) = P_{\mu_0}(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|) = ?$$

$P_{\mu_0}$ indicates that the probability is to be evaluated under $H_0 : \mu = \mu_0$. 
Case 1: \( n \) Large and \( \sigma \) Known

When \( n \) is large and \( \sigma \) is known the CLT gives us

\[
Z_n = \frac{\bar{X}_n - \mu_0}{\sigma / \sqrt{n}} \approx Z \sim \mathcal{N}(0, 1)
\]

and thus with \( z_n = (\bar{x}_n - \mu_0) / (\sigma / \sqrt{n}) = z.n \) we have

\[
P_{\mu_0}(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|) = P_{\mu_0} \left( \frac{|\bar{X}_n - \mu_0|}{\sigma / \sqrt{n}} \geq \frac{|\bar{x}_n - \mu_0|}{\sigma / \sqrt{n}} \right)
\]

\[
= P(|Z_n| \geq |z_n|) \approx P(|Z| \geq |z_n|) = 2\Phi(-|z_n|)
\]

\[
= 2 \times \text{pnorm}(-\text{abs}(z.n))
\]
A Situation with Known Variance

A natural example with known variance occurs in the binomial testing situation when using a normal approximation.

\[ X_1, \ldots, X_n \sim \text{Bernoulli}(p), \text{ then } \mu = EX_i = p \quad \text{and} \quad \sigma^2 = \text{var}X_i = p(1 - p). \]

Under the hypothesis \( H_0 : \mu = \mu_0 = p_0 \) the variance \( \sigma^2 = \sigma_0^2 = p_0(1 - p_0) \) is known and we can use for \( \bar{X}_n = (X_1 + \ldots + X_n)/n = Y_n/n \) the normal approximation

\[
Z_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\mu_0(1 - \mu_0)/n}} = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{Y_n - np_0}{\sqrt{np_0(1 - p_0)}} \approx Z \sim \mathcal{N}(0, 1)
\]
An Example

Example: Test $H_0 : \mu = 0.5$ against $H_1 : \mu \neq 0.5$ using $n = 2500$ trials with observed proportion of successes $\bar{x}_n = 1200/2500 = 0.48$.

Is this significant at level $\alpha = 0.05$, i.e., should we reject $H_0$ at this level?

\[
p(\bar{x}_n; \mu_0) = P_{\mu_0}(|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|) = P_{\mu_0}\left(\frac{|\bar{X}_n - \mu_0|}{\sqrt{\mu_0(1-\mu_0)/n}} \geq \frac{|\bar{x}_n - \mu_0|}{\sqrt{\mu_0(1-\mu_0)/n}}\right)
\]

\[
= P_{\mu_0}\left(|Z_n| \geq \frac{|0.48 - 0.5|}{\sqrt{0.5 \cdot 0.5/2500}}\right)
\]

\[
\approx P(|Z| \geq 2) = 2\Phi(-2) = 2 \times \text{pnorm}(-2) = 0.04550026 < 0.5
\]

The exact value via binomial calculation is $0.04768187$.

This is significant at level $\alpha = 0.05$, thus we reject $H_0$ (barely).
Case 2: Population Variance Unknown

For unknown population variance \( \sigma^2 \) (even under \( H_0 : \mu = \mu_0 \)) we estimate \( \sigma^2 \) using

\[
S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

and invoke the approximate normality of the test statistic

\[
T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \approx Z \sim \mathcal{N}(0, 1)
\]

As before, we calculate the significance probability of experiencing a value of \( |T_n| \) as extreme or more extreme than the observed one, i.e., \( |t_n| \)

\[
p(t_n; \mu_0) = P_{\mu_0} \left( \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \geq \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} \right)
\]

\[
= P_{\mu_0} (|T_n| \geq |t_n|) \approx P(|Z| \geq |t_n|) = 2 \Phi(-|t_n|)
\]
An engineering drawing calls out a nominal part dimension as 2.5 meters. To test whether the part supplier can, on average, produce this dimension, a sample of \( n = 50 \) produced parts is measured, with resulting values \( \bar{x}_n = 2.499\text{m} \) (off by 0.1cm) and \( s_n = 0.002\text{m} = 0.2\text{cm} \).

Should we reject the hypothesis \( H_0 : \mu = 2.5\text{m} \)?

\[
\frac{|2.499 - 2.5|}{0.002 / \sqrt{50}} = 3.535534
\]

\[
2 \times \text{pnorm}(-3.535534) = 0.0004069519
\]

a highly significant result. Reject \( H_0 \).

The value of \( s_n \) could be caused by the manufacturing and/or the measurement process. Had we gotten a smaller \( s_n \) (everything else the same) what might be our conclusion?
One-Sided Hypotheses and Alternatives

In our coin spinning example we simply tested whether the process was fair.

The real issue is whether whoever suggested coin spinning derived an advantage from some prior knowledge that it would favor Tails.

Thus it would be more appropriate to test

$$H_0 : p = P(\text{Heads}) \geq 0.5 \quad \text{against} \quad H_1 : p < 0.5$$

Under $H_1$ we have $P(\text{Tails}) > 0.5$, in favor of the suggester.

This is an example of a one-sided hypothesis and alternative. We don’t care if the evidence comes out in favor of $p \geq 0.5$ when we want to prove that the suggester had an advantage, $p < 0.5$. 
More generally, in terms of the population mean $\mu$ we have the following two canonical one-sided testing situations

$$H_0 : \mu \leq \mu_0 \quad \text{against} \quad H_1 : \mu > \mu_0$$

and

$$H_0 : \mu \geq \mu_0 \quad \text{against} \quad H_1 : \mu < \mu_0$$

where $\mu_0$ is a known value of separation belonging to $H_0$ in both cases.

The reason for this is that it is more practical for the calculation of significance probabilities.
We discuss the following three situations in the case of an unknown variance.

The known variance case is analogous: \( s_n \leftrightarrow \sigma \) and \( T_n \leftrightarrow Z_n \).

(a) \( H_0 : \mu = \mu_0 \) versus \( H_1 : \mu \neq \mu_0 \) \( |\bar{x}_n - \mu_0| \gg 0 \) speaks against \( H_0 \)

(b) \( H_0 : \mu \leq \mu_0 \) versus \( H_1 : \mu > \mu_0 \) \( \bar{x}_n - \mu_0 \gg 0 \) speaks against \( H_0 \)

(c) \( H_0 : \mu \geq \mu_0 \) versus \( H_1 : \mu < \mu_0 \) \( \bar{x}_n - \mu_0 \ll 0 \) speaks against \( H_0 \)

\( \gg \) means “very much larger than”

\( \ll \) means “very much smaller than,”

both without being specific about “very much.”
Tests in the One- and Two-Sided Cases

\[ |\bar{x}_n - \mu_0| \gg 0 \iff |t_n| = |\bar{x}_n - \mu_0|/(s_n/\sqrt{n}) \gg 0 \]  
with significance probability

\[ p_{(a)}(t_n; \mu_0) = P_{\mu_0}(|T_n| \geq |t_n|) \approx 2\Phi(-|t_n|) \]  
for large \( n \)

\[ \bar{x}_n - \mu_0 \gg 0 \iff t_n = (\bar{x}_n - \mu_0)/(s_n/\sqrt{n}) \gg 0 \]  
with significance probability

\[ p_{(b)}(t_n; \mu_0) = P_{\mu_0}(T_n \geq t_n) \approx 1 - \Phi(t_n) \]  
for large \( n \)

\[ \bar{x}_n - \mu_0 \ll 0 \iff t_n = (\bar{x}_n - \mu_0)/(s_n/\sqrt{n}) \ll 0 \]  
with significance probability

\[ p_{(c)}(t_n; \mu_0) = P_{\mu_0}(T_n \leq t_n) \approx \Phi(t_n) \]  
for large \( n \)

Small \( p_{(a)}(t_n; \mu_0), \ p_{(b)}(t_n; \mu_0), \ p_{(c)}(t_n; \mu_0) \) are reason to reject the respective \( H_0 \).
Suppose $\mu_0 = 20$ in our previous one- and two-sided testing problems and that with $n = 400$ we observe $\bar{x}_n = 21.82935$ and $s_n = 24.70037$.

Then

$$t_n = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{21.82935 - 20}{24.70037/\sqrt{400}} = 1.481233$$

with respective significance probabilities

$$p_{(a)}(t_n; \mu_0) = 2 \times \text{pnorm}(-1.481233) = 0.1385445$$
$$p_{(b)}(t_n; \mu_0) = 1 - \text{pnorm}(1.481233) = 0.06927225$$
$$p_{(c)}(t_n; \mu_0) = \text{pnorm}(1.481233) = 0.9307278$$

Note (and see Figure on next slide)

$$p_{(b)}(t_n; \mu_0) = p_{(a)}(t_n; \mu_0)/2 \quad \text{and} \quad p_{(b)}(t_n; \mu_0) + p_{(c)}(t_n; \mu_0) = 1$$
Significance Probabilities

(a)

(b)

(c)
Often it is debatable which of two one-sided hypotheses is appropriate.

Why was an experiment performed?

Who needs to be convinced of what?

Which error is more important than the other?
A group of parents wants the city to protect a school zone with speed humps. It is a 15 mph zone and it is agreed (by city & citizens) that an average speed > 15 mph should warrant speed humps, and < 15 mph should not.

The traffic was monitored and the average speed of $n = 150$ motorists was $\bar{x}_n = 15.3$ mph with $s_n = 2.5$ mph. How to set up the hypotheses?

**Parents’ perspective:** (risk of injury or life)
They want speed humps unless convinced otherwise, i.e., $H_0 : \mu \geq 15$ mph
They can live with a 1% chance of type I error, falsely say $\mu < 15$ mph.

**City’s perspective:** (financial risk)
Avoid speed humps (cost) unless convinced otherwise, i.e., $H_0 : \mu \leq 15$ mph
They allow for a 10% chance of type I error, falsely say $\mu > 15$ mph.
The observed test statistic is

\[ t_n = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}} = \frac{15.3 - 15}{2.5 / \sqrt{150}} = 1.469694 \]

Under the parents’ perspective there is no case against their hypothesis

\[ H_0 : \mu \geq 15 \text{ mph} \]. Any convincing evidence would require \( t_n \ll 0 \), but \( t_n = 1.47 \).

Under the city’s perspective \( t_n = 1.469694 \) speaks against \( H_0 : \mu \leq 15 \text{ mph} \).

How strongly?

\[ p_b(t_n; \mu_0) = 1 - \text{pnorm}(1.469694) = 0.07082232 < 0.10 \]

Since this falls below the city’s requirement (significance level \( \alpha = 0.10 \)),

the city has to follow through with installing the humps.

If the significance probability had exceeded \( \alpha = 0.10 \) we would have an impasse.

Is average speed relevant, given that 15 mph is a speed limit?

Maybe one should limit the 0.9-quantile of speed by 15 mph. (Binomial test)
An advertising campaign can claim increased mileage for a gasoline additive if it can be shown that the mileage increase $X$ (mileage with – without additive) on average exceeds 1 mpg. Test $H_0 : \mu = EX \leq 1$ mpg against $H_1 : \mu > 1$ mpg at significance level $\alpha = 0.05$, since we wish to “prove” $H_1$.

A corporation (with $\mu = 1.01$) tests $n = 900$ vehicles. It gets $\bar{x}_n = 1.01$ with $s_n = 0.1$.

$$t_n = \frac{1.01 - 1}{0.1/\sqrt{900}} = 3 \quad \text{with} \quad p(b)(t_n; \mu_0) \approx 1 - \text{pnorm}(3) = 0.001349898 \ll 0.05$$

An amateur mechanic (with $\mu = 1.21$) tests $n = 9$ cars, with $\bar{x}_n = 1.21$ & $s_n = 0.4$.

$$t_n = \frac{1.21 - 1}{0.4/\sqrt{9}} = 1.575 \quad \text{with} \quad p(b)(t_n; \mu_0) \approx 1 - \text{pnorm}(1.575) = 0.05762822 > 0.05$$

In case 1 we have statistical significance with little practical increase over 1 mpg.

In case 2 we have a 20% increase over 1 mpg, but no statistical significance, because of the small sample size and the higher $s_n$. ($\bar{x}_n = \mu$ is just illustrative)
Statistical significance is not the same as material or practical significance. They don’t preclude each other. Sometimes they occur together.

The first is based on probability calculations rooted in the statistical variability of experiments in conjunction with actual effects.

The second is an assessment of the actual effects in relation to some standard. This involves no randomness.
Set Estimation

Recall that in the binomial case we motivated set estimates as consisting of all those parameter values \( p_0 \) for which the hypothesis \( H_0 : p = p_0 \) is acceptable or plausible, i.e., not rejected at level \( \alpha \).

In the context of the mean \( \mu \) as our parameter of interest we view as set estimate of \( \mu \) all those values \( \mu_0 \) for which the hypothesis \( H_0 : \mu = \mu_0 \) is acceptable, i.e., not rejected at level \( \alpha \).

The construction of such set estimates can be implemented in the case of known \( \sigma \) and unknown \( \sigma \) (with \( S_n \) in its place), using the respective test statistics

\[
Z_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \quad \text{or} \quad T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}
\]

In each case \( n \) is assumed to be sufficiently large for the CLT to be effective.
Set Estimation: Unknown $\sigma$

Let $q = q_{1-\alpha/2}$ be the $(1 - \alpha/2)$-quantile of $N(0, 1)$, i.e.,

$$\Phi(q) = 1 - \alpha/2 \quad \text{or} \quad \Phi(-q) = \alpha/2$$

$\mu_0$ is acceptable whenever

$$p(a)(t_n; \mu_0) = P_{\mu_0}\left(\frac{|\bar{X}_n - \mu_0|}{S_n/\sqrt{n}} \geq \frac{|\bar{x}_n - \mu_0|}{s_n/\sqrt{n}}\right)$$

$$\approx 2\Phi(-|t_n|) > \alpha \iff -|t_n| > -q \iff |t_n| < q$$

$$|t_n| < q \iff |\bar{x}_n - \mu_0| < q \frac{s_n}{\sqrt{n}} \iff \mu_0 \in \left(\bar{x}_n - q \frac{s_n}{\sqrt{n}}, \bar{x}_n + q \frac{s_n}{\sqrt{n}}\right)$$

the righthand side interval serving as our set estimate, or plausibility interval, or confidence interval for the unknown $\mu$. (replace $s_n$ by $\sigma$ when $\sigma$ is known)

Different $\alpha$ give us different confidence intervals because of the $q$ factor.

$$(\bar{x}_n - q s_n/\sqrt{n}, \bar{x}_n + q s_n/\sqrt{n})$$ is called a $(1 - \alpha)$-level confidence interval for $\mu$. 

48
Coverage Probability

Treating the confidence intervals as random using the \((\bar{X}_n, S_n)\) notation

\[
I_1 = \left(\bar{X}_n - q \frac{S_n}{\sqrt{n}}, \bar{X}_n + q \frac{S_n}{\sqrt{n}}\right) \quad \text{and} \quad I_0 = \left(\bar{X}_n - q \frac{\sigma}{\sqrt{n}}, \bar{X}_n + q \frac{\sigma}{\sqrt{n}}\right)
\]

we get (for example) in the case of the \(I_1\)-interval (unknown \(\sigma\))

\[
P_{\mu_0}(\mu_0 \in I_1) = P_{\mu_0} \left(\bar{X}_n - q \frac{S_n}{\sqrt{n}} < \mu_0 < \bar{X}_n + q \frac{S_n}{\sqrt{n}}\right) = P_{\mu_0} \left(-q \frac{S_n}{\sqrt{n}} < \mu_0 - \bar{X}_n < q \frac{S_n}{\sqrt{n}}\right)
\]

\[
= P_{\mu_0} \left(|\mu_0 - \bar{X}_n| < q \frac{S_n}{\sqrt{n}}\right) = P_{\mu_0} \left(|\bar{X}_n - \mu_0| \frac{S_n}{\sqrt{n}} < q\right)
\]

\[
= P_{\mu_0}(|T_n| < q) \approx P(|Z| < q) = 1 - \alpha
\]

Since this coverage statement holds for any \(\mu_0\) we have in \(I_1\) a random interval that covers the unknown \(\mu\) with probability \(\approx 1 - \alpha\) (the confidence level).

Higher confidence level \(\iff\) higher \(q\) \(\iff\) wider interval.

The parameter \(\mu\) is unknown but not random in this representation.
Testing the Coverage Behavior

We obtain samples of size \( n = 30 \) \((n = 120)\) from the Uniform(0,1) population.

Without knowing from which population the sample originated, we compute a 90\% (95\%) confidence interval \((\bar{x}_n - q s_n/\sqrt{n}, \bar{x}_n + q s_n/\sqrt{n})\)

We repeat this 100 times and vertically plot the resulting intervals in sequence.

We mark the true mean \( \mu = 0.5 \) as a horizontal line

and indicate the percentage of interval misses.
Coverage Behavior of Confidence Intervals

10% misses

\[ 1 - \alpha = 0.9 \]

\[ n = 30 \]
Coverage Behavior of Confidence Intervals

- 14% misses
- $1 - \alpha = 0.9$
- n = 30
Coverage Behavior of Confidence Intervals

17% misses

$1 - \alpha = 0.9$

$n = 30$
Coverage Behavior of Confidence Intervals

6% misses

$1 - \alpha = 0.9$

$n = 30$
Coverage Behavior of Confidence Intervals 5

6% misses

$1 - \alpha = 0.95$

$n = 120$
Since the confidence set consists of all acceptable $\mu_0$, they can be used to test the hypothesis $H_0: \mu = \mu_0$. Tests and confidence sets are equivalent (duality).

Reject $H_0$ at level $\alpha$ whenever the $(1 - \alpha)$-level interval does not cover $\mu_0$.

The nature of confidence intervals has to be understood operationally.

For any given interval you never know whether you captured your target or not. You just know that you would have, for about $100(1 - \alpha)\%$ of the samples.

“Statistics means never having to say you’re certain.” (Myles Hollander)
Sample Size Planning

Assume that $\sigma$ is known.

For given confidence level $1 - \alpha$ and corresponding confidence factor $q = q_{1 - \alpha/2}$ the width $W$ of the interval

$$
\bar{x}_n \pm \frac{q\sigma}{\sqrt{n}} \quad \text{is} \quad W = \frac{2q\sigma}{\sqrt{n}}
$$

This allows us to achieve a specified width $W$ for proper choice of $n$, namely

$$
\sqrt{n} = \frac{2q\sigma}{W} \quad \text{or} \quad n = \left(\frac{2q\sigma}{W}\right)^2
$$

where we take the next higher integer for $n$.

When $\sigma$ is not known we need to take a guess at it or estimate it using $s_n$ as obtained from a prior sample. This is an approximate procedure.
One-Sided Confidence Intervals

You can invoke the testing and confidence set duality also for one-sided tests.

Then you get intervals of the form \((L, \infty)\) or \((-\infty, U)\),
i.e., you get lower or upper confidence bounds.

For example, you reject \(H_0 : \mu \leq \mu_0\) when the lower bound interval \((L, \infty)\)
does not overlap \((-\infty, \mu_0]\), i.e., whenever \(L \geq \mu_0\).

To be specific, consider testing \(H_0 : \mu \leq \mu_0\) against \(H_1 : \mu > \mu_0\).
At level \(\alpha\) we reject \(H_0\) whenever
\[
T_n = \left(\bar{X}_n - \mu_0\right) / \left(S_n / \sqrt{n}\right) \geq z_{1-\alpha},
\]
where \(\Phi(z_{1-\alpha}) = 1 - \alpha\). Conversely, \(H_0\) is acceptable whenever \(T_n < z_{1-\alpha}\), i.e.,
\[
\frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}} < z_{1-\alpha} \iff \bar{X}_n - \mu_0 < \frac{z_{1-\alpha} S_n}{\sqrt{n}} \iff L = \bar{X}_n - \frac{z_{1-\alpha} S_n}{\sqrt{n}} < \mu_0
\]
with coverage probability
\[
P_{\mu_0}(L < \mu_0) = P_{\mu_0}\left(\left(\bar{X}_n - \mu_0\right) / \left(S_n / \sqrt{n}\right) < z_{1-\alpha}\right) \approx 1 - \alpha.
\]
Gasoline Additives Revisited

For a 95% lower bound we need $z_{0.95} = \text{qnorm}(0.95) = 1.644854 \approx 1.645$.

For the corporation we get as observed lower bound $\ell_n$

$$\ell_n = \bar{x}_n - \frac{z_{1-\alpha} s_n}{\sqrt{n}} = 1.01 - 1.645 \cdot \frac{0.1}{\sqrt{900}} \approx 1.0045$$

i.e., a confidence interval of $(1.0045, \infty)$, which does not overlap the interval $(-\infty, 1]$ stated in the hypothesis $H_0 : \mu \leq 1$. Thus we reject $H_0$ at level $\alpha = 0.05$.

For the amateur mechanic the lower bound is

$$\ell_n = \bar{x}_n - \frac{z_{1-\alpha} s_n}{\sqrt{n}} = 1.21 - 1.645 \cdot \frac{0.4}{\sqrt{9}} \approx 0.9907$$

with confidence interval $(0.9907, \infty)$ which has a bit of overlap with $(-\infty, 1]$, i.e., we cannot reject $H_0$ at level $\alpha = 0.05$. 