Samples and Statistics

“The objective of statistical inference is to draw conclusions or make decisions about a population, based on a sample selected from the population.”

Inference is simplest when the sample is a random sample from the population: the sample values $X_1, X_2, \ldots, X_n$ are statistically independent and all have the same distribution.

That is not possible when sampling without replacement from a finite population; in that case, a random sample is one that is drawn in such a way that all $\binom{N}{n}$ possible samples have the same probability of being chosen.
It is not always possible or desirable to use a random sample.

For example, the successive values plotted in a control chart are rarely independent, because they are influenced by slow-changing properties of the system.

When we know, or suspect, that the sample was *not* a random sample, we should use appropriate methods.
Statistic

A **statistic** is a quantity that can be calculated from only the values in a sample.

Examples of statistics:

- **Sample mean:**

  \[
  \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i;
  \]

- **Sample standard deviation:**

  \[
  s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2};
  \]

A quantity like \( \bar{x} - \mu \) is *not* a statistic, because to calculate it we must know the value of the population parameter \( \mu \).
Sampling distribution

A statistic computed from a random sample itself a random variable, and has its own probability distribution.

The distribution of a statistic of a random sample is called its sampling distribution, to emphasize that we are dealing with a statistic and not a single observation.
Sampling from a normal distribution

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean $\mu$ and variance $\sigma^2$.

That is, $X_1, X_2, \ldots, X_n$ are independent, and each is distributed as $N(\mu, \sigma^2)$.

Then the sampling distribution of the sample mean $\bar{X}$ is $N(\mu, \sigma^2/n)$, or equivalently

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$
The sampling distribution of the sample variance is a scaled chi-square distribution:

$$\chi^2 = \frac{(n - 1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$ 

The $\chi^2$ distribution with $\nu$ degrees of freedom, here $n - 1$, is the Gamma distribution with shape parameter $r = \nu/2$ and rate parameter $\lambda = 1/2$. 
These sampling distributions are used to derive confidence intervals for $\mu$ and $\sigma^2$, respectively.

However, the confidence interval for $\mu$ requires that we know the value of $\sigma$; this is rarely the case.

When $\sigma$ is unknown, we use a third sampling result: the sampling distribution of

\[ T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \]

is Student’s $t$-distribution with $n - 1$ degrees of freedom.
Sampling from a Bernoulli distribution

Recall the notion of a sequence of independent trials, each resulting in success or failure, used to introduce the binomial distribution.

Let $X_i$ be the indicator of success at the $i^{th}$ trial:

$$X_i = \begin{cases} 
1 & \text{if the } i^{th} \text{ trial is a success;} \\
0 & \text{if the } i^{th} \text{ trial is a failure.}
\end{cases}$$

Each $X_i$ follows the Bernoulli distribution with parameter $p = P(X_i = 1)$. 
The number of successes in $n$ trials is $X = X_1 + X_2 + \cdots + X_n$, which follows the binomial distribution with parameters $n$ and $p$.

The sample mean $\bar{X} = X/n = \hat{p}$ also has a discrete distribution, most easily described in terms of the distribution of $X$; in particular

$$E(\bar{X}) = p$$

and

$$\text{Var}(\bar{X}) = p(1 - p)/n.$$ 

By the Central Limit Theorem, $\bar{X}$ is approximately normal, $N(p, p(1 - p)/n)$. 
Sampling from a Poisson distribution

If $X_1, X_2, \ldots, X_n$ are independent and each has the Poisson distribution with parameter $\lambda$, then $X = X_1 + X_2 + \cdots + X_n$ follows the Poisson distribution with parameter $n\lambda$.

The sample mean $\bar{X} = X/n = \hat{p}$ also has a discrete distribution, most easily described in terms of the distribution of $X$; in particular

$$E(\bar{X}) = \lambda$$

and

$$\text{Var}(\bar{X}) = \lambda/n.$$ 

By the Central Limit Theorem, $\bar{X}$ is approximately normal, $N(\lambda, \lambda/n)$. 
More generally, if $X_1, X_2, \ldots, X_n$ are independent and $X_i$ has the Poisson distribution with parameter $\lambda_i$, then $X = X_1 + X_2 + \cdots + X_n$ follows the Poisson distribution with parameter $\sum_{i=1}^{n} \lambda_i$. 
Point Estimation

In any of these sampling contexts, we need to make *inferences* about the parameter(s) of the corresponding model.

A **point estimator** of a parameter is a sample statistic that approximates the parameter.

As a statistic, it has a sampling distribution, with a mean and a variance.

The standard deviation of its sampling distribution is called its **standard error**.
If an estimator $\hat{\theta}$ of some parameter $\theta$ satisfies $E(\hat{\theta}) = \theta$, it is called **unbiased**.

In some situations, but not all, unbiased estimators are best.

The **mean squared error** of an estimator $\hat{\theta}$ of some parameter $\theta$ is

$$E[(\hat{\theta} - \theta)^2] = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

which for an unbiased $\hat{\theta}$ is just $\text{Var}(\hat{\theta})$.

In a random sample, the sample mean $\bar{X}$ and variance $s^2$ are always unbiased estimators of the population mean $\mu$ and variance $\sigma^2$, respectively, but $s$ is biased for $\sigma$. 
In some situations, the sample range, \( x(n) - x(1) \), has been used to construct an estimator of the population standard deviation \( \sigma \) because it requires little computation.

This construction is critically dependent on the assumption that the data are normally distributed; for any other distribution, the relationship between the range and the standard deviation is different.
Inference for a Single Sample

Inferences about some parameter may be made using:
- a point estimator;
- an interval estimator;
- a hypothesis test.
Mean of a normal population

Point estimator

The usual point estimator of $\mu$ is the unbiased $\bar{X}$.

The sampling distribution of $\bar{X}$ is $N(\mu, \sigma^2/n)$, so its standard error is $\sigma/\sqrt{n}$.

When $\sigma$ is unknown, we replace it by $s$ to get the estimated standard error $s/\sqrt{n}$.
Interval estimator

The usual interval estimator is a **confidence interval**, derived from the distribution of \( Z \) (when \( \sigma \) is known) or \( T \) (when \( \sigma \) is unknown).

- **Known \( \sigma \):**
  \[
  \bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}
  \]

- **Unknown \( \sigma \):**
  \[
  \bar{X} \pm t_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}
  \]

In each case, the interval contains \( \mu \) with probability \( 1 - \alpha \), and is called a \( 100(1 - \alpha)\% \) confidence interval.

The **confidence level** \( 100(1 - \alpha)\% \) is often 95%, but sometimes 99% is preferred.
Example: Computer response time

- Assumed to be normal with $\sigma = 8$ msec.
- Mean of 25 measured response times is $\bar{x} = 79.25$ msec, the point estimate.
- Standard error is $8/\sqrt{25} = 1.6$ msec.
- The 95% confidence interval is

$$79.25 \pm 1.96 \times 1.6 = (76.11, 82.39) \text{ msec}$$

Confidence interpretation: the statement

$$76.11 \leq \mu \leq 82.39$$

was made using a procedure that has a 95% chance of being correct.
Example: Viscosity of rubberized asphalt

- Assumed to be normal.
- Mean of 15 measured viscosities is $\bar{x} = 3210.73$ cP, the point estimate of $\mu$.
- Standard deviation of 15 measured viscosities is $s = 117.61$ cP, the point estimate of $\sigma$.
- Estimated standard error of $\bar{x}$ is $s / \sqrt{15} = 30.367$ cP.
- The 95% confidence interval is

$$3210.73 \pm 2.145 \times 30.367 = (3145.60, 3275.86) \text{ cP}$$

Again: we have 95% confidence in these statements because they are made in a procedure that has a 95% chance of producing correct statements.
Hypothesis testing

Point and interval estimates are not guided by any distinguished value of the parameter.

When some particular value is of special interest, a **hypothesis test** may be appropriate.

**Example: Computer response time**

Does the mean response time exceed 75 msec?
The **null hypothesis** is that the performance is acceptable:

\[ H_0 : \mu \leq \mu_0 = 75. \]

The **alternate hypothesis** is that the performance is bad:

\[ H_1 : \mu > \mu_0. \]

The **test statistic** is

\[ z_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{79.25 - 75}{8/\sqrt{25}} = 2.66. \]

We reject \( H_0 \) if \( z_{\text{obs}} \) is too large.
Two types of error:

Type I error (false positive): Rejecting $H_0$ when it is true.

Type II error (false negative): Failing to reject $H_0$ when it is false.

We usually specify $\alpha = P(\text{Type I error})$; often $\alpha = 0.05$, sometimes $\alpha = 0.01$.

To achieve a Type I error rate $\alpha = 0.05$, we reject $H_0$ if $z_{\text{obs}} > z_\alpha = z_{0.05} = 1.645$.

So in this case we reject $H_0$. That is, the data are inconsistent with the hypothesis that the mean response time is at most 75 msec.
Example: Viscosity of rubberized asphalt

In this case the nominal viscosity is 3200 cP, and deviation in either direction is bad:

\[ H_0 : \mu = \mu_0 = 3200, \quad H_1 : \mu \neq \mu_0. \]

Now \( \sigma \) is unknown, so we cannot use \( z_{obs} \). We replace \( \sigma \) by \( s \), so the test statistic is

\[
t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
\]

\[
= \frac{3210.73 - 3200}{117.61/\sqrt{15}}
\]

\[
= 0.35.
\]
In this case, we reject $H_0$ if the magnitude of $t_{obs}$ is too large.

To achieve $\alpha = 0.05$, we reject $H_0$ if

$$|t_{obs}| > t_{\alpha/2,n-1} = t_{0.025,14} = 2.145.$$

So for these data we do not reject $H_0$. That is, the observed data are consistent with the hypothesis that the mean viscosity is the nominal value 3200 cP.

**Language**

Never state that we *accept* the null hypothesis. Failing to reject $H_0$ means only that it is a reasonable approximation in the light of the observed data.
P-Value

The result of a hypothesis test (reject $H_0$, or do not reject $H_0$) does not convey much information.

The test statistic might be very close to the critical value, or very far from it.

We could test at several $\alpha$ levels: in the computer response time example, we reject $H_0$ when $\alpha = 0.05$, and also when $\alpha = 0.01$ ($z_{0.01} = 2.326$), but not when $\alpha = 0.001$ ($z_{0.001} = 3.090$).

The **P-value** is the smallest $\alpha$ for which we reject $H_0$.

A small $P$-value is strong evidence against the null hypothesis.
For tests about $\mu$ with known $\sigma$, $P$ is:

- $2[1 - \Phi(|z_{\text{obs}}|)]$ for a two-tailed test, $\mu = \mu_0$ versus $\mu \neq \mu_0$;
- $1 - \Phi(z_{\text{obs}})$ for an upper-tailed test, $\mu \leq \mu_0$ versus $\mu > \mu_0$;
- $\Phi(z_{\text{obs}})$ for a lower-tailed test, $\mu \geq \mu_0$ versus $\mu < \mu_0$;

Example: Computer response time

Here $z_{\text{obs}} = 2.66$ in an upper-tailed test, so $P = 1 - \Phi(2.66) = 0.0039$.

So $H_0$ would be rejected in any test with $\alpha \geq 0.0039$. 
When \( \sigma \) is unknown, replace \( z_{\text{obs}} \) with \( t_{\text{obs}} \) and use the cdf of the \( t \)-distribution instead of \( \Phi(\cdot) \).

**Example: Viscosity of rubberized asphalt**

Here \( t_{\text{obs}} = 0.35 \) in a two-tailed test, so

\[
P = 2[1 - F_{t,14}(0.35)] = 0.73.
\]

So \( H_0 \) would not be rejected in a test with any of the usual \( \alpha \) levels (\( \alpha = 0.1 \) is the largest value usually considered).
Power of a test

The probability of a Type II error (failing to reject a false null hypothesis) is usually denoted $\beta$.

The probability of correctly rejecting a false $H_0$ is the **power**, $1 - \beta$. The power of a test depends on how far $\mu$ is from the value(s) in the null hypothesis.

Power is sometimes used to decide sample size.
Example: Computer response time

Suppose the mean response time is actually 80 msec. The test statistic

\[ Z_{\text{obs}} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \]

now has expected value

\[ E(Z_{\text{obs}}) = \frac{\mu - \mu_0}{\sigma / \sqrt{n}} \]

\[ = \frac{80 - 75}{8 / \sqrt{25}} \]

\[ = 3.125 \]

so \( Z_{\text{obs}} \sim N(3.125, 1) \).
The power of the test is

\[ P(Z_{obs} > 1.645) = P(Z_{obs} - 3.125 > 1.645 - 3.125) = 1 - \Phi(-1.48) = 0.93. \]

That is, there is a 93% chance of rejecting \( H_0 : \mu \leq 75 \) when the mean is actually 80 msec.
Variance of a normal population

Point estimator

In a random sample $X_1, X_2, \ldots, X_n$ from any population with variance $\sigma^2$, the sample variance

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

is an unbiased estimator of $\sigma^2$. 

Interval estimator

Instead of providing the standard error of $S^2$, we usually provide a confidence interval.

Recall that when the population is normal,

$$\chi^2 = \frac{(n - 1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$
Now

\[ 1 - \alpha = P\left(\chi_{1-\alpha/2,n-1}^2 \leq \chi^2 \leq \chi_{\alpha/2,n-1}^2\right) \]

\[ = P \left[ \chi_{1-\alpha/2,n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2,n-1}^2 \right] \]

\[ = P \left[ \frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2,n-1}^2} \right]. \]

So

\[ \frac{(n-1)S^2}{\chi_{1-\alpha/2,n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\alpha/2,n-1}^2} \]

is a 100(1 - \alpha)% confidence interval for \( \sigma^2 \).
Example: Viscosity of rubberized asphalt

The sample standard deviation is $s = 117.61$ cP so the 95% confidence interval for $\sigma$ is

$$117.61 \sqrt{\frac{14}{\chi^2_{0.025,14}}} \leq \sigma \leq 117.61 \sqrt{\frac{14}{\chi^2_{0.975,14}}}$$

or

$$86.11 \leq \sigma \leq 185.48.$$ 

The interval is quite wide: a sample of size $n = 15$ does not give precise information about $\sigma$. 
We may be more concerned about high values of variance than low values, because high variance means less precise measurements.

The 100(1 − \(\alpha\))% upper confidence bound for \(\sigma^2\) is

\[
\sigma^2 \leq \frac{(n - 1)S^2}{\chi^2_{1-\alpha,n-1}}.
\]

For the viscosity data, this gives \(\sigma \leq 171.67\) cP.
Hypothesis test

To test \( H_0 : \sigma^2 = \sigma_0^2 \) against the two-sided alternative \( H_1 : \sigma^2 \neq \sigma_0^2 \), we similarly use the test statistic

\[
\chi^2_{obs} = \frac{(n - 1)s^2}{\sigma_0^2}.
\]

For a test with Type I error rate \( \alpha \), reject \( H_0 \) if \( \chi^2_{obs} \) is too far in either tail:

\[
\chi^2_{obs} > \chi^2_{\alpha/2,n-1} \quad \text{or} \quad \chi^2_{obs} < \chi^2_{1-\alpha/2,n-1}
\]

Alternatively, the \( P \)-value is

\[
P = 2 \min \left[ F_{\chi^2,n-1}(\chi^2_{obs}), 1 - F_{\chi^2,n-1}(\chi^2_{obs}) \right].
\]
Example: Viscosity of rubberized asphalt

Test \( H_0 : \sigma = 100 \) versus \( H_1 : \sigma \neq 100 \). The sample standard deviation is \( s = 117.61 \) so

\[
\chi^2_{\text{obs}} = \frac{14 \times 117.61^2}{100^2} = 19.365.
\]

\( F_{\chi^2,14}(19.365) = 0.8485 \), so \( P = 0.303 \): the data are consistent with the null value \( \sigma = 100 \).
Tests about variance are often one-sided; for example, $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$.

In this case, reject $H_0$ only if $\chi^2_{\text{obj}}$ is too large:

$$\chi_{\text{obs}}^2 > \chi^2_{\alpha,n-1}$$

Alternatively, the $P$-value is

$$P = 1 - F_{\chi^2,n-1}(\chi_{\text{obs}}^2).$$
Inference for a Population Proportion

The context now is a random sample $X_1, X_2, \ldots, X_n$ from the Bernoulli distribution with probability $p$, or equivalently an observation $X$ from the binomial distribution with parameters $n$ and $p$.

**Point estimator**

The sample fraction $\bar{X} = X/n = \hat{p}$ is an unbiased estimator of $p$.

The standard error of $\hat{p}$ is $\sqrt{p(1-p)/n}$, and the estimated standard error is $\sqrt{\hat{p}(1-\hat{p})/n}$. 
Interval estimator

The simplest interval estimator is the approximate confidence interval based on the normal approximation to the binomial distribution:

\[ \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}. \]

Because it is based on an approximation, the coverage probability of this interval estimator may differ from the nominal 100(1 - \(\alpha\))%.

Many alternatives have been proposed and studied.
Hypothesis test

To test the hypothesis $H_0 : p = p_0$ against the two-sided alternative $H_1 : p \neq p_0$, use the test statistic

$$Z = \begin{cases} 
\frac{(X + 0.5) - np_0}{\sqrt{np_0(1 - p_0)}} & \text{if } X < np_0 \\
\frac{(X - 0.5) - np_0}{\sqrt{np_0(1 - p_0)}} & \text{if } X > np_0
\end{cases}$$

where $X = n\hat{p}$ is the number of “successes”.

This statistic is approximately $N(0, 1)$, so you carry out a test with Type I error rate $\alpha$ or compute a $P$-value in the usual way.