UQ, STAT2201, 2017,
Lecture 7.
Unit 7 – Single Sample Inference.

Setup: A sample $x_1, \ldots, x_n$ (collected values).

Model: An i.i.d. sequence of random variables, $X_1, \ldots, X_n$.

Parameter at question: The population mean, $\mu = E[X_i]$.

Point estimate: $\overline{x}$ (described by the random variable $\overline{X}$).

Goal: Devise hypothesis tests and confidence intervals for μ.

Distinguish between the two cases:

- Unrealistic (but simpler): The population variance, σ^2, is known.
- More realistic: The variance is not known and estimated by the sample variance, s^2.

For very small samples, the results we present are valid only if the population is normally distributed.

But for non-small samples (e.g. $n > 20$, although there isn't a clear rule), the central limit theorem provides a good approximation and the results are approximately correct.

Testing Hypotheses on the Mean, Variance Known (Z-Tests)

Model:	$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$	with μ unknown but σ^2 known.
Null hypothesis:	$H_0 : \mu = \mu_0$.	
Test statistic:	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \qquad Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$	

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\left[1 - \Phi(\lvert z \rvert)\right]$	$z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$P = 1 - \Phi(z)$	$z > z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$P = \Phi(z)$	$z < z_{\alpha}$

For $H_1 : \mu \neq \mu_0$, a procedure identical to the preceding fixed significance level test is:

$$\text{Reject } H_0 : \mu = \mu_0 \qquad \text{if either} \qquad \bar{x} < a \text{ or } \bar{x} > b$$

where

$$a = \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{and} \qquad b = \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Compare with the confidence interval formula:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

If H_0 is not true and H_1 holds with a specific value of $\mu = \mu_1$, then it is possible to compute the probability of type II error, β.

In the (very realistic) case where σ^2 is not known, but rather estimated by S^2, we would like to replace the test statistic, Z, above with,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

but in general, T no longer follows a Normal distribution.

Under $H_0 : \mu = \mu_0$, and for moderate or large samples (e.g. $n > 100$) this statistic is approximately Normally distributed just like above. In this case, the procedures above work well.

But for smaller samples, the distribution of T is no longer Normally distributed. Nevertheless, it follows a well known and very famous distribution of classical statistics: **The Student-t Distribution**.

The probability density function of a Student-t Distribution with a parameter k, referred to as **degrees of freedom**, is,

$$f(x) = \frac{\Gamma\left[(k+1)/2\right]}{\sqrt{\pi k}\Gamma(k/2)} \cdot \frac{1}{\left[\left(x^2/k\right)+1\right]^{(k+1)/2}} \qquad -\infty < x < \infty,$$

where $\Gamma(\cdot)$ is the Gamma-function. It is a symmetric distribution about 0 and as $k \to \infty$ it approaches a standard Normal distribution.

Why is the t-distribution so useful in (small sample) elementary statistics?

Claim: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a Normal distribution with mean μ and variance σ^2. The random variable, T has a t distribution with $n-1$ degrees of freedom.

Knowing the distribution of T (and noticing it depends on the sample size, n), allows to construct hypothesis tests and confidence intervals when σ^2 is not known.

The construction is analogous to the Z-tests and confidence intervals.

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1-\alpha)\%$ **confidence interval** on μ is given by

$$\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of the t distribution with $n-1$ degrees of freedom.

A related concept is a $100(1-\alpha)\%$ **prediction interval** (PI) on a single future observation from a normal distribution is given by

$$\bar{x} - t_{1-\alpha/2,n-1} s \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{1-\alpha/2,n-1} s \sqrt{1 + \frac{1}{n}}.$$

This is the range where we expect the $n+1$ observation to be, after observing n observations and computing $\overline{x}$ and s.

Testing Hypotheses on the Mean, Variance Unknown (T-Tests)

Model:	$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$	with both μ and σ^2 unknown
Null hypothesis:	$H_0 : \mu = \mu_0$.	
Test statistic:	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$.	

Alternative Hypotheses	P**-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\left[1 - F_{n-1}\left(\lvert t\rvert\right)\right]$	$t > t_{1-\alpha/2,n-1}$ or $t < t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$P = 1 - F_{n-1}\left(t\right)$	$t > t_{1-\alpha,n-1}$
$H_1 : \mu < \mu_0$	$P = F_{n-1}\left(t\right)$	$t < t_{\alpha,n-1}$

In the P-value calculation, $F_{n-1}(\cdot)$ denotes the CDF of the t-distribution with $n-1$ degrees of freedom.

As opposed to $\Phi(\cdot)$, the CDF of t is not tabulated in standard tables. So to calculate P-values, we use software (or make educated guesses using quantiles).