STAT 509 – Section 3.6: Sampling Distributions

Definition: <u>Parameter</u> = a number that characterizes a population (example: population mean μ) – it's typically <u>unknown</u>.

<u>Statistic</u> = a number that characterizes a sample (example: sample mean $\overline{Y}$) – we can calculate it from our sample data.

$\overline{Y}$ =

We use the sample mean $\overline{Y}$ to estimate the population mean μ.
Suppose we take a sample and calculate $\overline{Y}$.
Will $\overline{Y}$ equal μ? Will $\overline{Y}$ be close to μ?
Suppose we take another sample and get another $\overline{Y}$.
Will it be same as first $\overline{Y}$? Will it be close to first $\overline{Y}$?

• What if we took <u>many repeated samples</u> (of the same size) from the same population, and each time, calculated the sample mean?
What would that set of $\overline{Y}$ values look like?

The <u>sampling distribution</u> of a statistic is the distribution of values of the statistic in all possible samples (of the same size) from the same population.

Consider the sampling distribution of the sample mean $\overline{Y}$ when we take samples of size *n* from a population with mean μ and variance σ^2.

Picture:

The sampling distribution of $\overline{Y}$ has mean μ and standard deviation $\sigma / \sqrt{n}$.

Notation:

<u>Central Limit Theorem</u>

We have determined the *center* and the *spread* of the sampling distribution of $\overline{Y}$. What is the *shape* of its sampling distribution?

Case I: *If the distribution of the original data is normal,* the sampling distribution of $\overline{Y}$ is normal. (This is true no matter what the sample size is.)

Case II: <u>Central Limit Theorem</u>: If we take a random sample (of size *n*) from any population with mean μ and standard deviation σ, the sampling distribution of $\overline{Y}$ is *approximately normal, if the sample size is large*.

How large does *n* have to be?
<u>One rule of thumb</u>: If $n \geq 30$, we can apply the CLT result.

Depends on the shape of the population distribution:

• If the data come from a distribution that is nearly normal, sample size need not be very large to invoke CLT.
• If the data come from a distribution that is far from normal, sample size must be very large to invoke CLT.

Pictures:

As *n* gets larger, the closer the sampling distribution looks to a normal distribution.

• Checking how close data are to being normally distributed can be done via <u>normal probability plots</u>.

• Normal probability (Q-Q) plots plot the ordered data values against corresponding N(0,1) quantiles:

<u>Ordered data</u>: $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$
<u>Normal Quantiles</u>: z-values with area $P_{(i)}$ to their left, for $i=1,\ldots,n$,

where $P_{(i)} = (i - 0.5) / n$

• In practice this is always plotted on a computer.

R code:

```
> qqnorm(mydata)
```

• If the plotted points fall in roughly a straight line, the assumption that the data are nearly normally distributed is reasonable.

• If the plotted points form a curve or an S-shape, then the data are not close to normal, and we need quite a large sample size to apply the CLT.

• Similar types of Q-Q plot can be used to check whether data may come from other specific distributions.

<u>Why is the CLT important</u>? Because when $\bar{Y}$ is (approximately) normally distributed, we can answer probability questions about the sample mean.

Standardizing values of $\bar{Y}$:

If $\bar{Y}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n}$, then

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution.

Example: The time between adjacent accidents in an industrial plant follows an exponential distribution with an average of 700 days. What is the probability that the average time between 49 pairs of adjacent accidents will be greater than 900 days?

Other Sampling Distributions

In practice, the population standard deviation σ is typically unknown.

We estimate σ with the sample standard deviation *s*,

where the sample variance $s^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

But the quantity $\dfrac{\bar{Y} - \mu}{s/\sqrt{n}}$ **does not have a standard normal distribution.**

Its sampling distribution is as follows:

• If the data come from a normal population, then the statistic $T = \dfrac{\bar{Y} - \mu}{s/\sqrt{n}}$ **has a t-distribution ("Student's t") with *n* – 1 degrees of freedom (the parameter of the t-distribution).**

• The t-distribution resembles the standard normal (symmetric, mound-shaped, centered at zero) but it is more spread out.

• The fewer the degrees of freedom, the more spread out the t-distribution is.

• As the d.f. increase, the t-distribution gets closer to the standard normal.

Picture:

Table 2 gives values of the t-distribution with specific areas to the left of these values.

Example: The nominal power produced by a student-designed internal combustion engine is 100 hp. The student team that designed the engine conducted 10 tests to determine the actual power. The data were:

```
97.9 100.8 102.0  97.0 100.8  97.9 100.1
91.9  98.1  99.9
```

Note for these data, n = 10, $\overline{Y}$ = 98.64, s = 2.864.

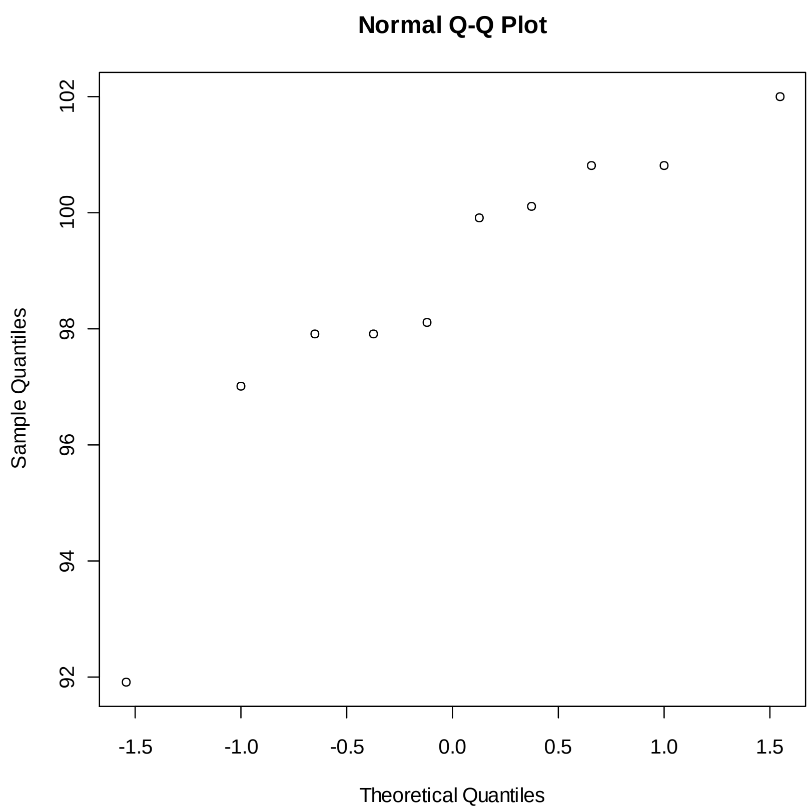
Assuming the data came from a normal distribution, what is the probability of getting a sample mean of 98.64 hp or less if the true mean is actually 100 hp?

Picture:

R code:

```
> pt(-1.502, df=9)
[1] 0.08367136
```

Is the normality assumption reasonable?



The χ^2 (Chi-square) Distribution

Suppose our sample (of size *n*) comes from a normal population with mean μ and standard deviation σ.

Then $\frac{(n-1)s^2}{\sigma^2}$ **has a χ^2 distribution with *n* – 1 degrees of freedom.**

- **The χ^2 distribution takes on positive values.**
- **It is skewed to the right.**
- **It is less skewed for higher degrees of freedom.**
- **The mean of a χ^2 distribution with *n* – 1 degrees of freedom is *n* – 1 and the variance is 2(*n* – 1).**

<u>Fact</u>: If we add the squares of *n* independent standard normal r.v.'s, the resulting sum has a χ^2_n distribution.

Note that $\frac{(n-1)s^2}{\sigma^2}$ **=**

We sacrifice 1 d.f. by estimating μ with $\overline{Y}$, so it is χ^2_{n-1}.

Table 3 gives values of a χ^2 r.v. with specific areas to the left of those values.

Examples:

For χ^2 with 6 d.f., area to the left of __________ is .10.

For χ^2 with 6 d.f., area to the left of __________ is .95.

For χ^2 with 20 d.f., area to the left of _________ is .90.

The F Distribution

The quantity $\dfrac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}$ **where the two** χ^2 **r.v.'s are independent, has an F-distribution with** $n_1 - 1$ **"numerator degrees of freedom" and** $n_2 - 1$ **denominator degrees of freedom.**

So, if we have independent samples (of sizes n_1 **and** n_2**) from <u>two</u> normal populations, note:**

has an F-distribution with ($n_1 - 1$, $n_2 - 1$**) d.f.**

Table 4 (p. 580) gives values of F r.v. with area .10 to the right.
Table 4 (p. 582) gives values of F r.v. with area .05 to the right.
Table 4 (p. 584) gives values of F r.v. with area .01 to the right.

Verify:

For F with (3, 9) d.f., 2.81 has area 0.10 to right.

For F with (15, 13) d.f., 3.82 has area 0.01 to right.

• These sampling distributions will be important in many inferential procedures we will learn.