MAT 222 Day One Review Spring 2017

What is Statistics?

Science of data.

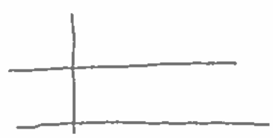
Note: data ≠ numbers. Have both quantitative & qualitative

These studies include:

- Collection
- Analysis
- Organization/Presentation

To study data, we look at its distribution – what values it takes and how often it takes the values – which we often plot. Some specific distributions are:

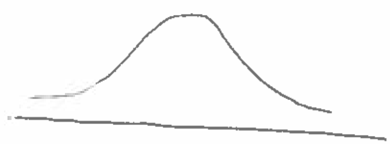
- Uniform Distribution:



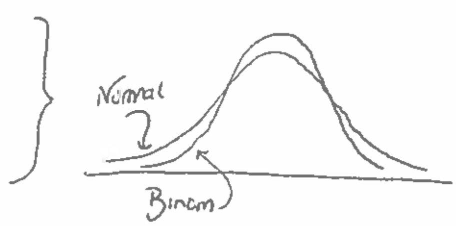
e.g., cards, dice, heads/tails

* Distribution ≠ Density curve
1) Nonneg
2) Area 1
3) Median half area
4) Mean Balance point

- Normal Distribution:



- Binomial Distribution:



We can see skewness from the plot of a distribution. As an example, look at the normal distribution:

Symmetric
Mean = Median

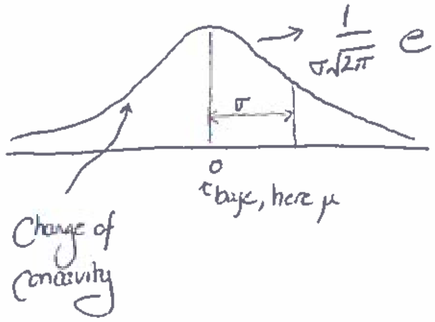
right skew
Median > Mean

left skew
Mean < Median

When looking at distributions, we are often most interested in examining the following:

- Median
- Mean * Which are resistant to outliers?
- Standard Deviation

One of the most common distributions which occurs in Statistics is the normal distribution (also called the Gaussian distribution or standard normal curve). We often write $N(\mu, \sigma)$ for the normal distribution with mean μ and standard deviation σ.



$\rightarrow \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$

$Z = \frac{x-\mu}{\sigma}$ } take to 'standard' normal

68-95-99.7 Rules

How do we begin discussing the probability of specific events coming from a distribution? One way is using the standardized value (also called the z-score):

$Z_x = \frac{x-\mu}{\sigma}$ ← distance (directed) from mean } # standard dev. from mean

↑ why?

* Gives 'uniqueness' of value
* Can be used to find prob of events.

$Z_x = 1.960 \nleftarrow 95\%$

* $Z_x = 1.960 \mapsto 95\%$

However, we are not often looking at an entire population. Rather we are looking at data from a sample of a given population. What is true for the *entire* population need *not* be true about the sample. If we take a SRS of size n from a population with mean μ and standard deviation σ, the mean and standard deviation of the samples are:

$\mu_S = \mu$ ← Why?

$\sigma_S = \frac{\sigma}{\sqrt{n}}$ ← Why?

In fact, if the sample size is large the sampling distribution is approximately normal with distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. [This is precisely the Central Limit Theorem.] We can even use normal curves (under certain conditions) to approximate samples from a binomial $B(n, p)$ distribution: given a SRS of size n from a large population having success p, then

$X \approx N(np, \sqrt{np(1-p)})$

Why?

$\hat{p} = N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Why?

* $np \geq 10$

$n(1-p) \geq 10$

But often we do not know the mean of the population we are examining. However since we know the underlying distribution, we can use information from a SRS to give estimations *with error* for the mean of the underlying population. This is precisely the notion of confidence intervals.

estimate ± margin of error

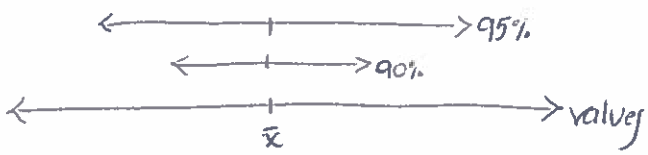
$\bar{x} \pm$ m.o.e.

$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

* Give normal curve on pp. 359 to get m.o.e.

$\#\text{s.d.} \cdot \text{s.d.}$

Graphically, we can represent and compare confidence intervals as follows:



See graphs on pp. 363

Why does it look like this?

To reduce the margin of error, we can:

- lower confidence C
- Reduce σ
- Choose larger sample

In fact, this last method can tell us how to choose our sample size:

$$n = \left(\frac{z^* \sigma}{m} \right)^2 \leftarrow \text{How?}$$

* In practice, still not enough
* Only applies in specific instances
* Error can still be larger than predicted.

The idea of confidence intervals allows us to give a method of testing the truth of a hypothesis against observed data. This is the idea of significance testing.

1 { H_0: def
  H_a: def

2 { test stat. def

3 { p-value (stat. sign. @ level α)

4 { Conclusion

* Why 5% sig? 1931 "The Design of Exp.", Fisher said '1 out of 20 might be reasonable.'

$\mu = 1000$

$\sigma = 40$

$n = 100$

$\mu_s = 1007$

1) $H_0: \mu = 1000$

$H_a: \mu > 1000$

2) $z_{1007} = \frac{1007-1000}{40/\sqrt{100}} = 1.925$

3) $1.925 \mapsto 0.9728 \to 97.28\%$

2.72%

Of course, statistical inference must be implemented carefully. There are many things to consider, especially with the experimental design:

* Cannot correct flaws in design
* Plan study to verify test your plan. Needs high prob. in detecting effect.
* Don't apply inference to data randomly - need confidence model.
* Cannot apply data then suggest hypothesis.

Of course, we could accept/reject H_0 when H_0 is true/false. This results in our statistical inferences being right/wrong. We need a way of measuring the likelihood of this occurring.

Decision based on sample.		Truth about the population: H_0 true	Truth about the population: H_a true
	Reject H_0	Type I	✓
	Accept H_0	✓	Type II

We need a way of measuring and discussing these errors.

Power: Prob level α sig. test reject H_0 when H_a is true.
Power to detect alternative.

1) State H_0, H_a
2) Find $\bar{x}$ needed to reject H_0
3) Calc. prob. of observing these values when alt. true.

Z score value with $\mu = \mu_s$
all σ are σ_s

6 month program to inc. bone min. content
1% change important. $\sigma = 2$ for % change.
Is 25 subjects enough?

1) $H_0: \mu = 0$
$H_a: \mu > 0$

2) Reject @ 0.05: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{\bar{x} - 0}{2/\sqrt{25}} \geq 1.645$

$\bar{x} \geq 0.658$ (0.05 chance when mean is 0)

3) To detect 1% change
$P(\bar{x} \geq 0.658 \text{ when } \mu = 1) = P\left(\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \dfrac{0.658 - 1}{2/\sqrt{25}}\right)$
$= P(z \geq -0.855)$
$= 0.80$

To increase the power, one could:

- Increase α : 5% less evidence than 1%
- Consider alt. farther from H_0 : 2 vs. 2.2 v.s. 2 vs. 200000
- Increase sample size
- Decrease sigma

Now in terms of Type I and Type II errors, we have:

Type I: prob. level α

Type II: 1 - power